**IMDB Movie Analysis**

**Project Description**

Investigating the factors that influence a movie's success on IMDb is a crucial endeavour. By defining success in terms of high IMDb ratings, we can delve into various aspects that contribute to a film's popularity and acclaim. Movie producers, directors, and investors stand to gain valuable insights from such an analysis, enabling them to make informed decisions for their future projects.

To address this problem, we can explore a range of features related to movies, including:

* Genre: Does the genre significantly impact IMDb ratings? Certain genres may resonate more with audiences, leading to higher ratings.
* Cast and Crew: The involvement of renowned actors, directors, or writers could positively influence a movie's reception.
* Budget and Box Office Performance: Is there a correlation between a film's budget and its IMDb rating? Additionally, how well a movie performs at the box office might affect its perceived success.
* Release Date: Does the timing of a movie's release (e.g., during holidays or blockbuster seasons) impact its IMDb rating?
* User Reviews and Ratings: Analysing user-generated reviews and ratings can provide insights into audience preferences.
* Runtime: Is there an optimal movie length that tends to receive better ratings?
* Awards and Nominations: Recognition from prestigious awards can boost a film's reputation.

By exploring these factors, we can uncover patterns and trends that contribute to a movie's success, ultimately benefiting the entertainment industry as a whole.

**Approach**

# **Data Cleaning:**

This initial step focuses on preprocessing the data to make it suitable for analysis, which in general is called Data Cleaning. It encompasses tasks such as handling missing values, removing duplicates, dropping redundant columns, converting data types (if needed), and potentially performing feature engineering. By ensuring data quality, we set the foundation for meaningful insights in subsequent stages of analysis.

* **The Dataset details are:**

1. Number of Data-Points: 5,043
2. Number of Features: 28

* **Column Details:**

1. color: Movie is Colored or Black and White

2. director\_name: Name of the movie’s director

3. num\_critic\_for\_reviews : Number of reviews by film critics

4. duration: Duration of the movie

5. director\_facebook\_likes: Facebook Likes of the director

6. actor\_3\_facebook\_likes : Facebook Likes of one of the actors

7. actor\_2\_name: Name of one of the actors

8. actor\_1\_facebook\_likes: Facebook Likes of one of the actors

9. gross: Gross collection of the movie

10. genres: Genrs of the movie

11. actor\_1\_name: Name of one of the actors

12. movie\_title: Name of the movie

13. num\_voted\_users: Number of users voted for the movie

14. cast\_total\_facebook\_likes: Movie cast’s total facebook likes

15. actor\_3\_name: Name of one of the actors

16. facenumber\_in\_poster: Number of faces in the movie’s poster

17. plot\_keywords: Some keywords from plot of the movie

18. movie\_imdb\_link: IMDB link of the movie

19. num\_user\_for\_reviews: Number of users who reviewed the movie

20. language: Original language of the movie

21. country: Country of origin of the movie

22. content\_rating: Content rating of the movie(Certification tag)

23. budget: Budget of the movie

24. title\_year: Year in which the movie was released

25. actor\_2\_facebook\_likes: Facebook Likes of one of the actors

26. imdb\_score: IMDB Score of the movie

27. aspect\_ratio: Aspect ratio in which the movie was made

28. movie\_facebook\_likes: Facebook likes of the movie

* **Dropping Redundant columns:**

Dropping columns during data analysis is a common operation that involves removing one or more columns from a dataset. If certain columns are irrelevant to your analysis or research question, consider dropping them. Irrelevant features can introduce noise and unnecessarily complicate your model. Here, out of 28 columns, only 12 are essential for our analysis. Hence, the rest of the redundant columns are dropped to make the data concise and relevant to the analysis.

* **Find and replace null values:**

Column ‘director\_name’ had a few empty cells or NULL values, which were found and replaced with the word ‘unknown’, where as the rest of the columns with NULL values were left as it is, but a filter has been added to better sort the sheet.

* **Handling Outliers:**

In the dataset, we observe outliers in the ‘budget,’ ‘gross,’ and ‘profit’ columns. These outliers represent valid numerical values but are exceptionally high. Instead of using averages (mean), it’s advisable to rely on medians or quantiles during our analysis. This approach helps mitigate the impact of extreme values and provides a more robust understanding of the data

* **Handling Errors and invalid values:**

All the invalid values present in the ‘movie\_title’ and ‘director\_name’ columns have been removed for better results during analysis.

# **Data Analysis:**

In this phase, your goal is to explore the data comprehensively. You’ll investigate the relationships between various variables.

Specifically, consider examining:

1. Correlation: Explore how IMDb ratings correlate with factors such as genre, director, budget, and other relevant features.
2. Release Year: Does the year of release impact ratings? Analyse whether older or newer movies tend to receive higher ratings.
3. Cast and Crew: Investigate whether specific actors, directors, or writers influence movie ratings.
4. Budget vs. Ratings: Explore whether higher budgets lead to better IMDb ratings.

By delving into these aspects, you’ll gain insights that can inform decision-making and potentially enhance future movie projects.

# **The “Five Whys Approach” for Root Cause Analysis:**

The Five Whys approach is a powerful tool used in Root Cause Analysis (RCA), its purpose is to identify the fundamental cause of a problem by repeatedly asking “Why?” in a sequence. Rather than merely addressing surface-level symptoms, this method aims to understand and tackle deeper issues. Start with a clear problem statement. Ask “why” to determine the initial cause of the problem. Continue asking “why” at least five times, probing deeper with each iteration. The goal is to get to the heart of the matter, beyond superficial explanations.

The 5 Whys Technique provides a systematic approach to problem-solving, helps prevent treating symptoms without understanding the underlying issues, enables organizations to make informed decisions based on true causes and is applicable across various sectors, from manufacturing to services. For example, if you observe that movies with higher budgets tend to receive higher ratings, you can use this approach to uncover the underlying reasons behind this correlation. By iteratively questioning the cause, you can reveal valuable insights that go beyond surface-level observations.

Using the “Five Whys Approach”, we have to find insights for:

1. **Common Genres and Descriptive Statistics:**

First, we’ll determine the most common genres of movies in the dataset. For each genre, we’ll calculate descriptive statistics (mean, median, mode, range, variance, and standard deviation) of the IMDb scores. This will give us insights into the overall rating distribution within each genre.

1. **Movie Durations and IMDb Scores:**

We’ll analyse the distribution of movie durations. This will help us understand the typical length of movies in the dataset. Next, we’ll explore the relationship between movie duration and IMDb score. Does longer or shorter duration correlate with higher ratings?

1. **Common Languages and Impact on IMDb Scores:**

We’ll determine the most common languages used in movies. Using descriptive statistics, we’ll assess how language impacts IMDb scores. Are certain languages associated with better ratings?

1. **Top Directors and Contribution to Success:**

Identifying the top directors based on their average IMDb score is crucial. We’ll analyse their contribution to movie success using percentile calculations. Are these directors consistently delivering high-rated films?

1. **Movie Budgets, Gross Earnings, and Profit Margin:**

We’ll explore the correlation between movie budgets and gross earnings. Additionally, we’ll identify movies with the highest profit margin (earnings relative to budget).

**Tech-Stack Used**

## **Microsoft Excel 2019**

Several features of Microsoft Excel were particularly useful for this project: data filtering for sorting and identifying patterns, pivot tables for summarizing large datasets, graphical tools for visualizing trends, formulas and functions for complex calculations, data cleaning tools for preprocessing accuracy, and conditional formatting for highlighting key insights. These collectively made Excel a powerful tool for analysing and visualizing the hiring process data.

**Insights**

# **Movie Genre Analysis:**

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

For movie genre analysis, creating of a new set of genre columns, specifically showing the genre distribution is supposed to be completed using excel COUNTIF function to count the number of movies for each genre. Top genres are calculate using pivot table and aggregation on IMDB score using average, then sorted by average. IMDB score movies with user vote greater than 25k has been considered for chart depiction.

* Creating Genre Column:

Start by creating a new genre column. Column represents a specific genre (e.g., Action, Drama, Comedy). Use the Excel COUNTIF function to count the number of movies associated with each genre. This will give you the genre distribution across the dataset.

Calculate mean, median, mode, standard deviation of the IMDB scores as well.

* Top Genres Calculation:

Calculate the top genres based on IMDb scores. Create a pivot table and aggregate IMDb scores using the average function (mean) for each genre. Sort the genres by their average IMDb score in descending order.

## **Insights:**

* **Number of movies by genre:**

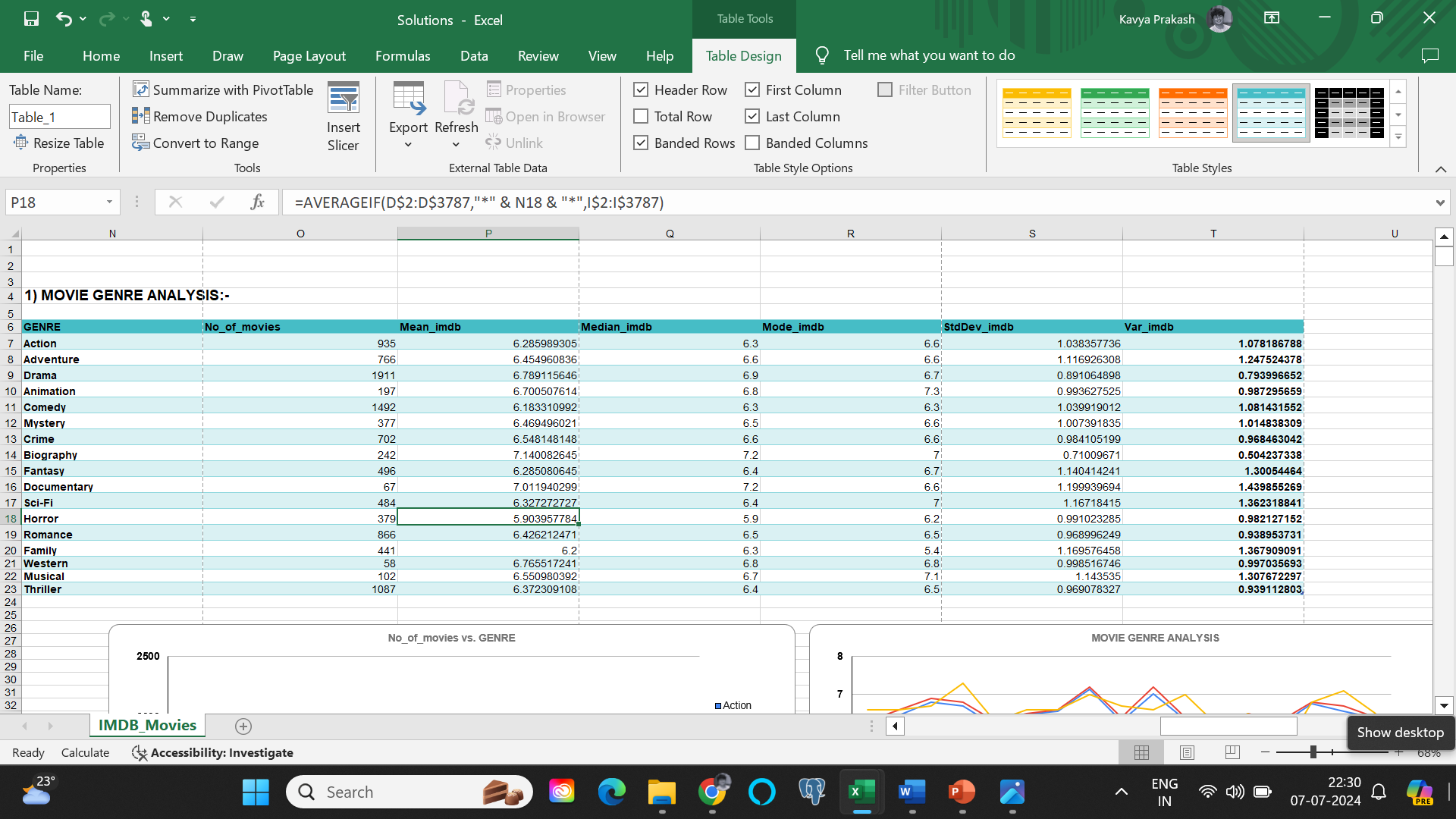
Among these genres, “Drama,” “Comedy,” and “Thriller” stand out with the highest number of movies made.

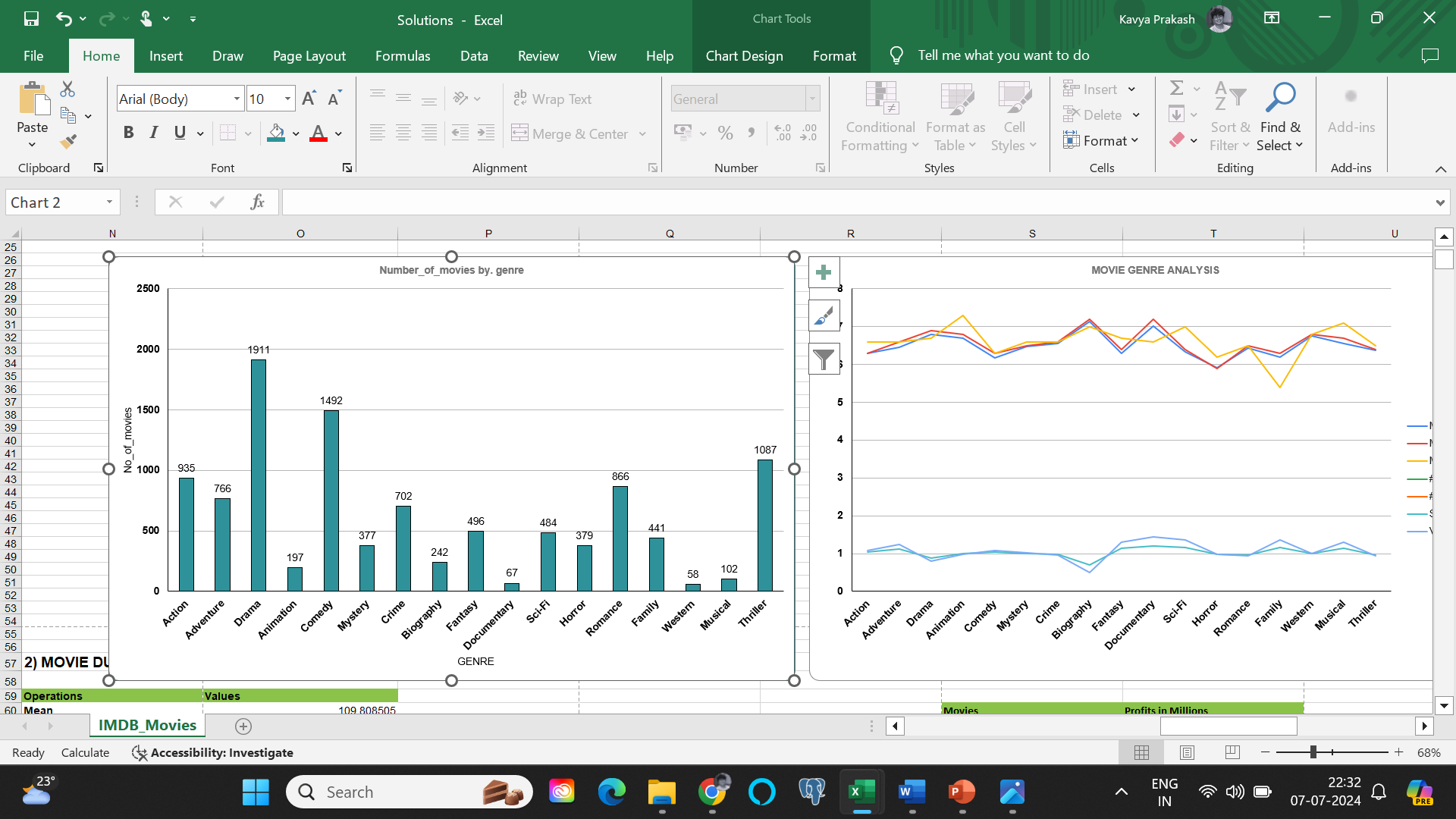
* **Recommendations for Genre Selection:**

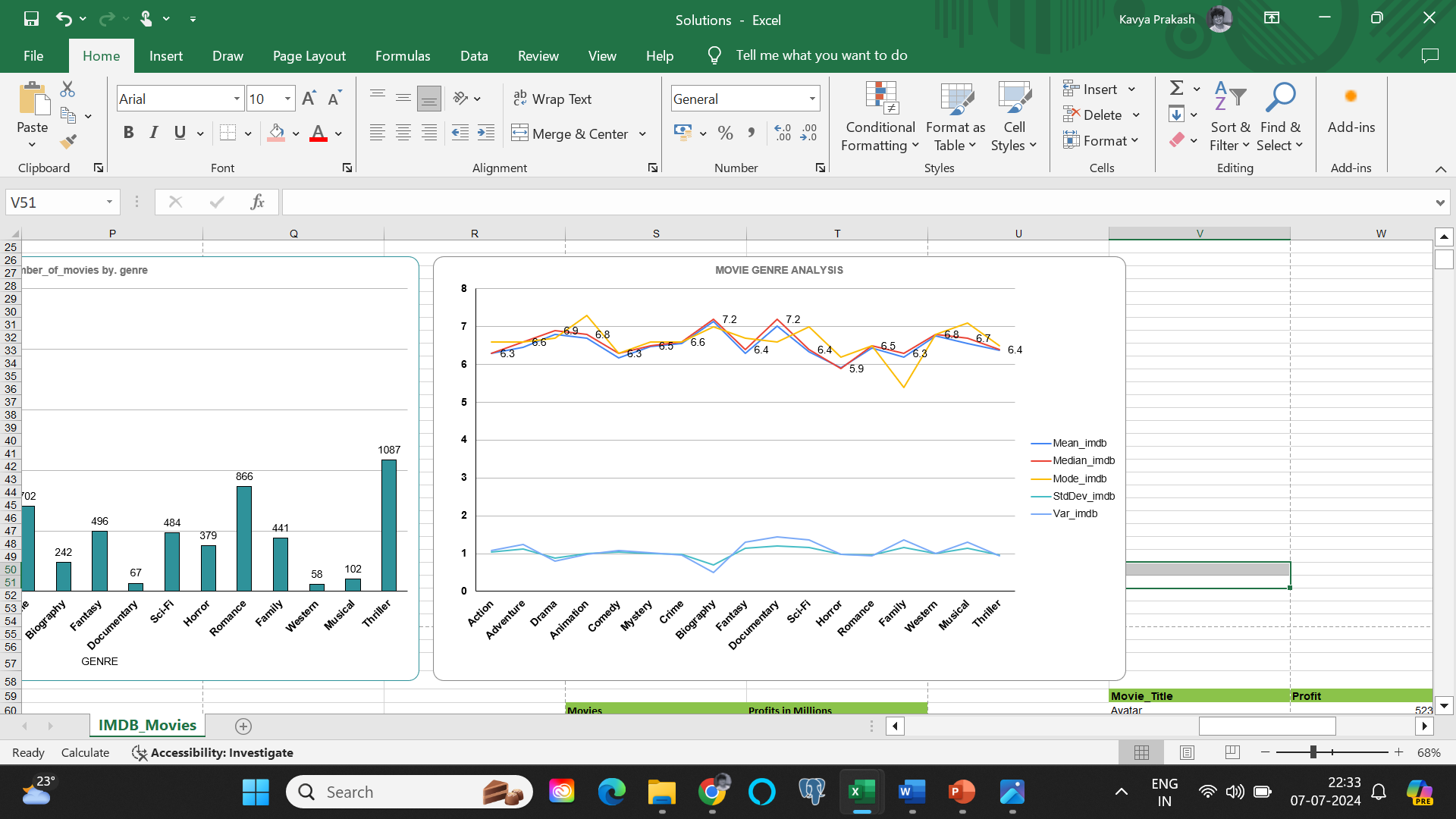
Among these genres, “Biography,” “History,” and “War” stand out with the highest mean IMDb ratings. For the company’s next movies, considering these genres could lead to well-received films. These genres have consistently delivered quality content, making them promising choices.

* **Key Takeaways:**

In summary, focusing on genres like biography, historical dramas, and war films could be a strategic move for the company’s future projects







# **Movie Duration Analysis:**

Analyse the distribution of movie durations and its impact on the IMDB score. Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

* **Descriptive Statistics for Movie Durations:**

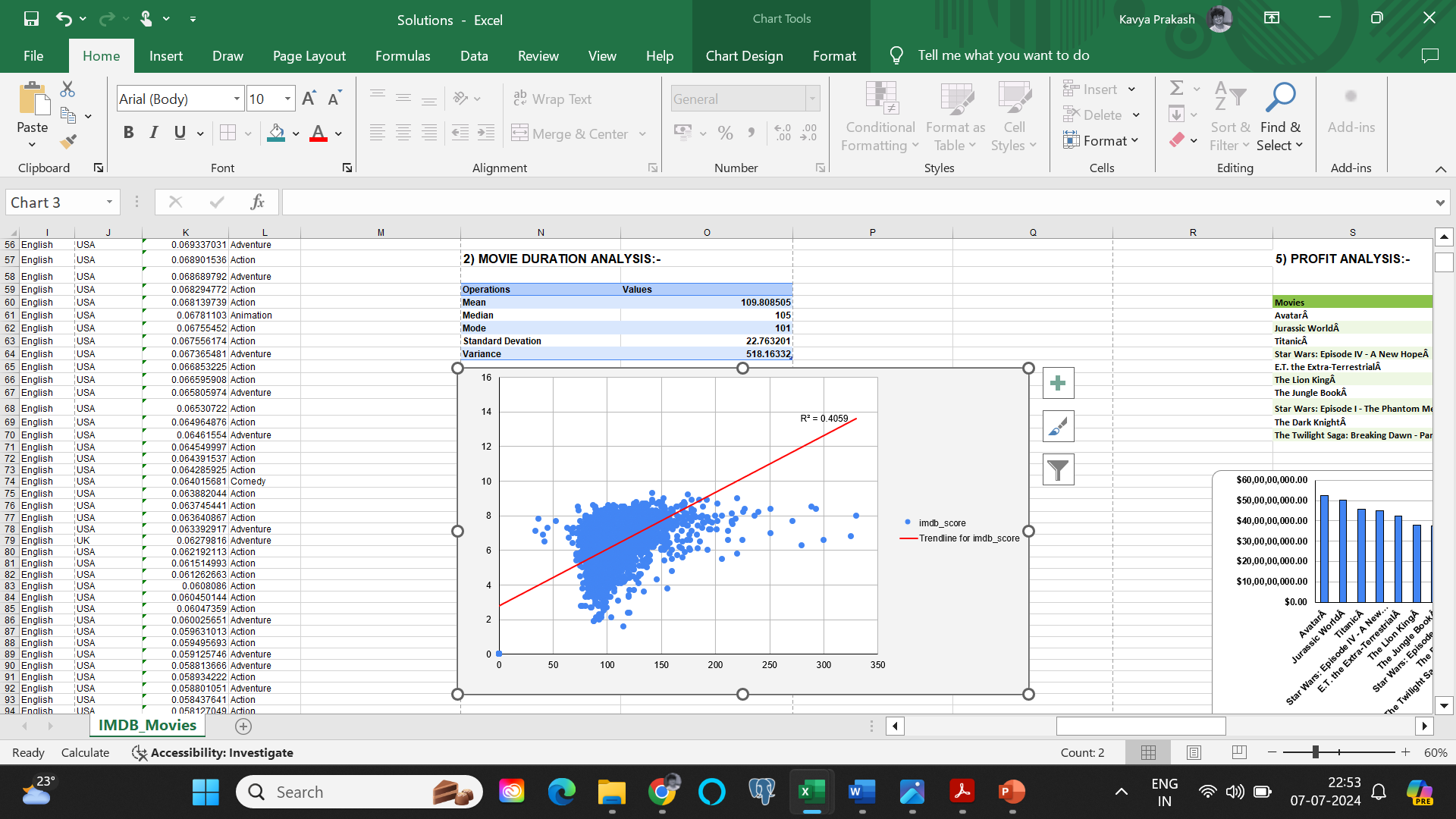
Calculate the mean, median, and standard deviation for movie durations. Excel’s functions like AVERAGE, MEDIAN, and STDEV will help you compute these statistics.

* **Scatter Plot Visualization:**

Create a scatter plot to visualize the relationship between movie duration and IMDb score. Plot movie durations on the x-axis and IMDb scores on the y-axis.

* **Trendline Assessment:**

Add a trendline to the scatter plot. The trendline will help assess the direction (positive or negative) and strength of the relationship between duration and IMDb score.



## **Insights:**

The average duration of a movie is 109 minutes. When examining the relationship between movie duration and IMDb score, the trendline shows an upward slope. The coefficient of determination (R-squared) for this trendline is 0.131

# **Language Analysis:**

Examine the distribution of movies based on their language. Determine the most common languages used in movies and analyse their impact on the IMDB score using descriptive statistics.

* **Counting Movies by Language:**

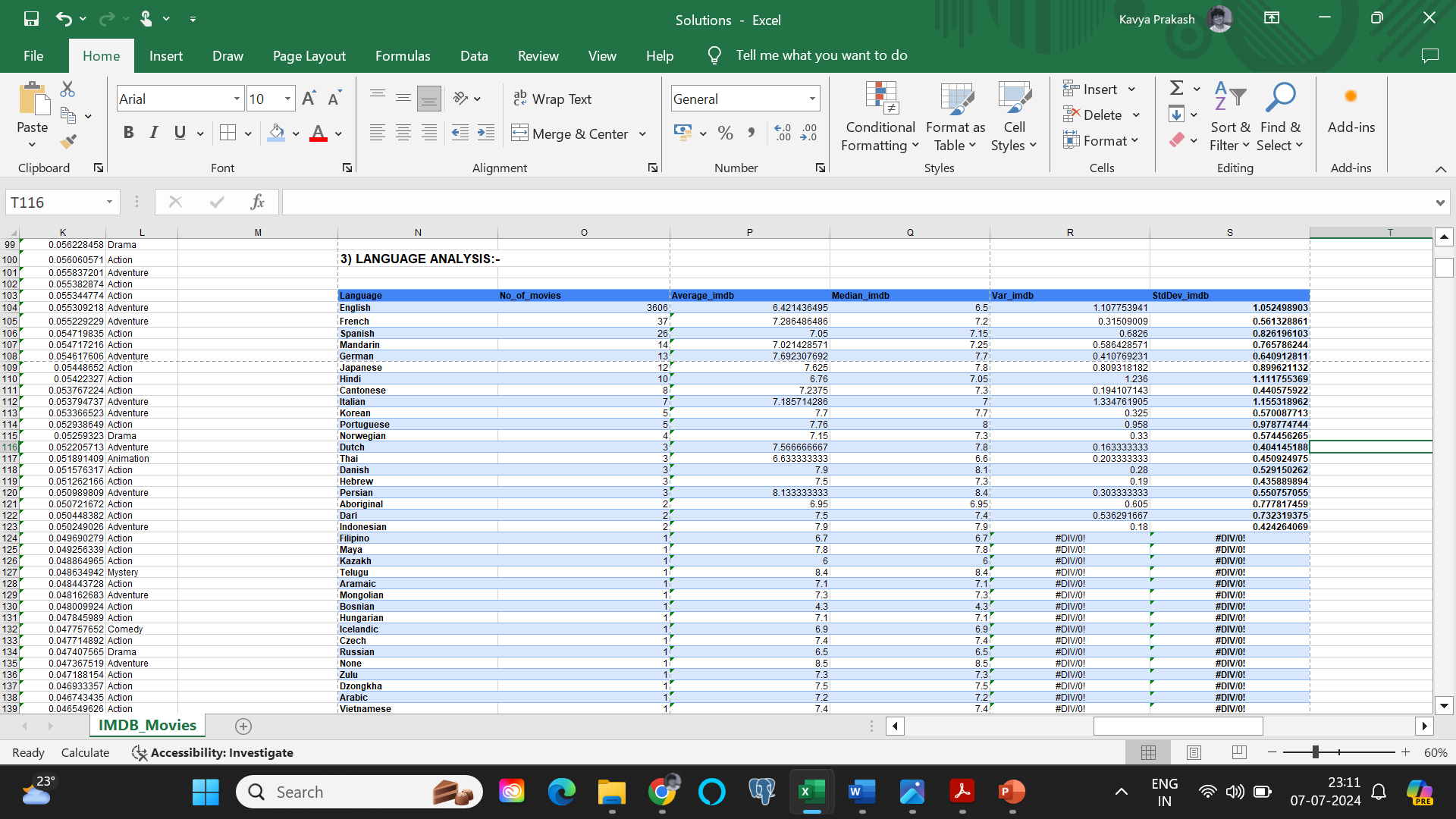
Utilize Excel’s COUNTIF function to count the number of movies for each language. This step will give you an overview of how movies are distributed across different languages.

* **Descriptive Statistics for IMDb Scores:**

Calculate the mean, median, and standard deviation of IMDb scores for each language. These statistics will help you understand the central tendency and variability of ratings within each language group.

* **Impact of Language on Movie Ratings:**

Compare the descriptive statistics across languages. Assess whether certain languages tend to have higher or lower IMDb scores. This analysis will provide insights into the impact of language on movie ratings.



## **Insights:**

Among the languages commonly used in movies, English, French, Spanish, Mandarin, and German stand out(highest movies made in these languages). Surprisingly, Telugu(8.4) and Persian(8.14) films tend to achieve the highest average IMDb scores.

# **Director Analysis:**

Influence of directors on movie ratings. Identify the top directors based on their average IMDB score and analyse their contribution to the success of movies using percentile calculations.

* **Calculate Average IMDb Score for Each Director:**

First, gather data on movies, including their directors and IMDb scores. Calculate the AVERAGE IMDb score for each director by taking the mean of the scores of all movies directed by them.

* **Identify Directors with High IMDb Scores using percentage comparision:**

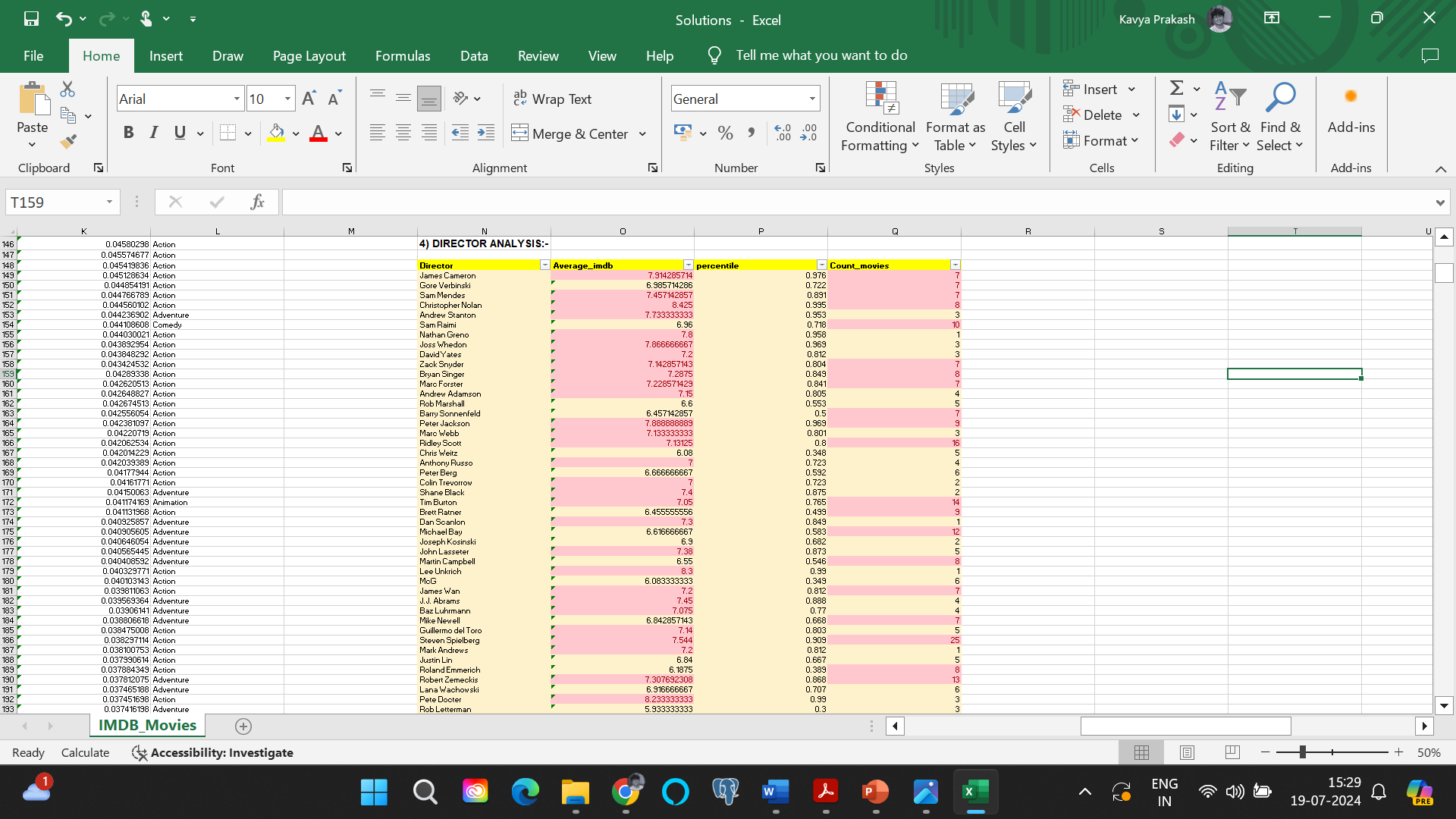
Use the PERCENTILE function in Excel to find the IMDb score that corresponds to a specific percentile. Identify directors whose average IMDb scores are above this percentile. These are the top directors based on IMDb ratings.

* **Additionally, identified the top 10% of Average and filtered to find the top directors:**

The average values of IMDb scores have been sorted and the top 10% Average vales and count of movies have been highlighted with a red fill marking.

## **Insights:**

After calculation of Average and percentile, we can see that Tony Kaye, Charles Chaplin, Alfred Hitchcock, Ron Fricke, Damien Chazelle, Majid Majidi, Sergio Leone, Christopher Nolan, SS Rajamouli and Richard Marquand as the top 10 directors with average IMDb score greater than or equal to 8.4. Additionally, top 10% of Averages and count of movies are highlighted, for better understanding of the data.



# **Budget Analysis:**

Explore the relationship between movie budgets and their financial success. Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

To explore the relationship between movie budgets and their financial success, Calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function. Scatter plots are created using charts and then, top movies are created using sorting by profit.

* Correlation Coefficient Calculation:

Use Excel’s CORREL function to calculate the correlation coefficient between movie budgets and gross earnings. This coefficient will indicate the strength and direction of the linear relationship between these two variables.

* Profit Margin Calculation:

Calculate the profit margin for each movie by subtracting the budget from the gross earnings. The profit margin represents the financial success of a movie (positive margin indicates profit, negative margin indicates loss).

* Identifying Top Movies:

Use Excel’s MAX function to find the highest profit margin among all movies. Create scatter plots using charts to visualize the relationship between budgets and gross earnings.

Sort the movies by profit margin to identify the top-performing ones.

## **Insights:**

* **Profitable Movies:**

“Avatar” stands out as the most profitable movie, closely followed by “Jurassic World.”

Both of these movies have substantial budgets but managed to generate impressive gross earnings.

* **Surprising Success Stories:**

“Star Wars” and “E. T. the Extra-Terrestrial” are remarkable examples.

Despite having relatively low budgets, they achieved high profit margins.

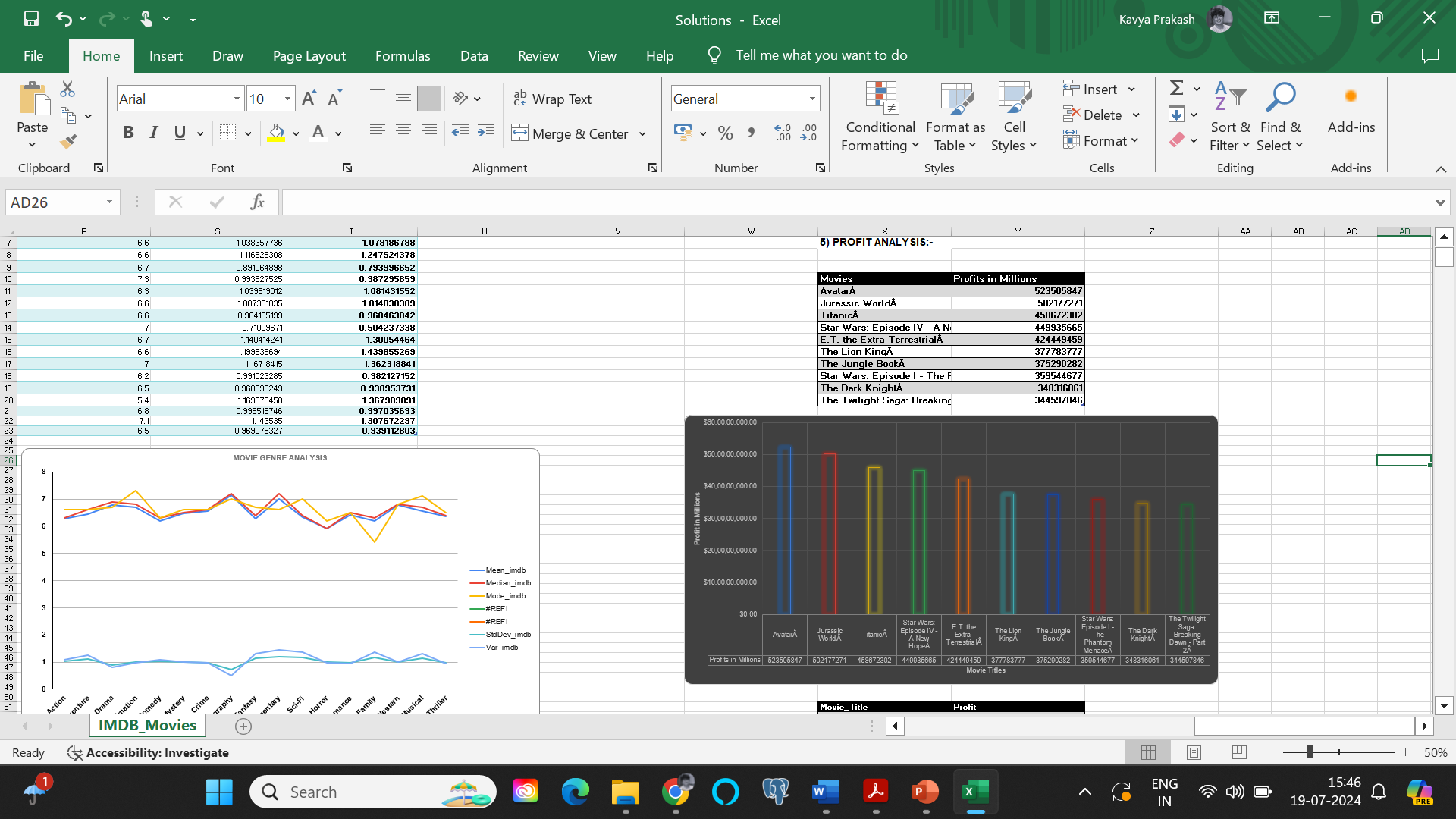
Their success defies the conventional notion that high budgets always lead to greater profits.

* **Budget-Profit Relationship:**

Overall, there isn’t a strong linear correlation between budget and profit.

While some high-budget movies do result in significant losses, others (like “Avatar” and “Jurassic World”) yield substantial profits.

The key takeaway: Success depends on factors beyond budget alone. In summary, the movie industry is full of surprises, and financial success isn’t solely determined by budget size



**Result**

Ideally, a film’s IMDb rating is influenced by various factors, including its duration, genre, language, budget, and director. However, the intricate interplay among these variables highlights the multifaceted nature of audience preferences and the ever-evolving landscape of cinematic success. Filmmakers and movie enthusiasts can gain valuable insights from this analysis, aiding their decision-making and deepening their understanding of the film industry dynamics. This project helped me asses the IMDB data, that provided innumerable insights into the intricate factors that influence the success rate of a movie on the IMDB website.

**Data sheet**

* [Project5 data solutions.xlsx](Project5%20data%20solutions.xlsx)

Or

* <https://docs.google.com/spreadsheets/d/177bgUWepPjcRRdbSTtH-yKU8aTRLRGk8/edit?usp=sharing&ouid=117438581175883142153&rtpof=true&sd=true>
* Opening the data sheet on Google sheets changes a few important design aspects. Hence, please open the Excel sheet on Microsoft Excel.