

# IMDB MOVIE ANALYTICS

## Trainity Project 5 (Final Project -1) – Advanced Statistics [Descriptive Analysis]

---

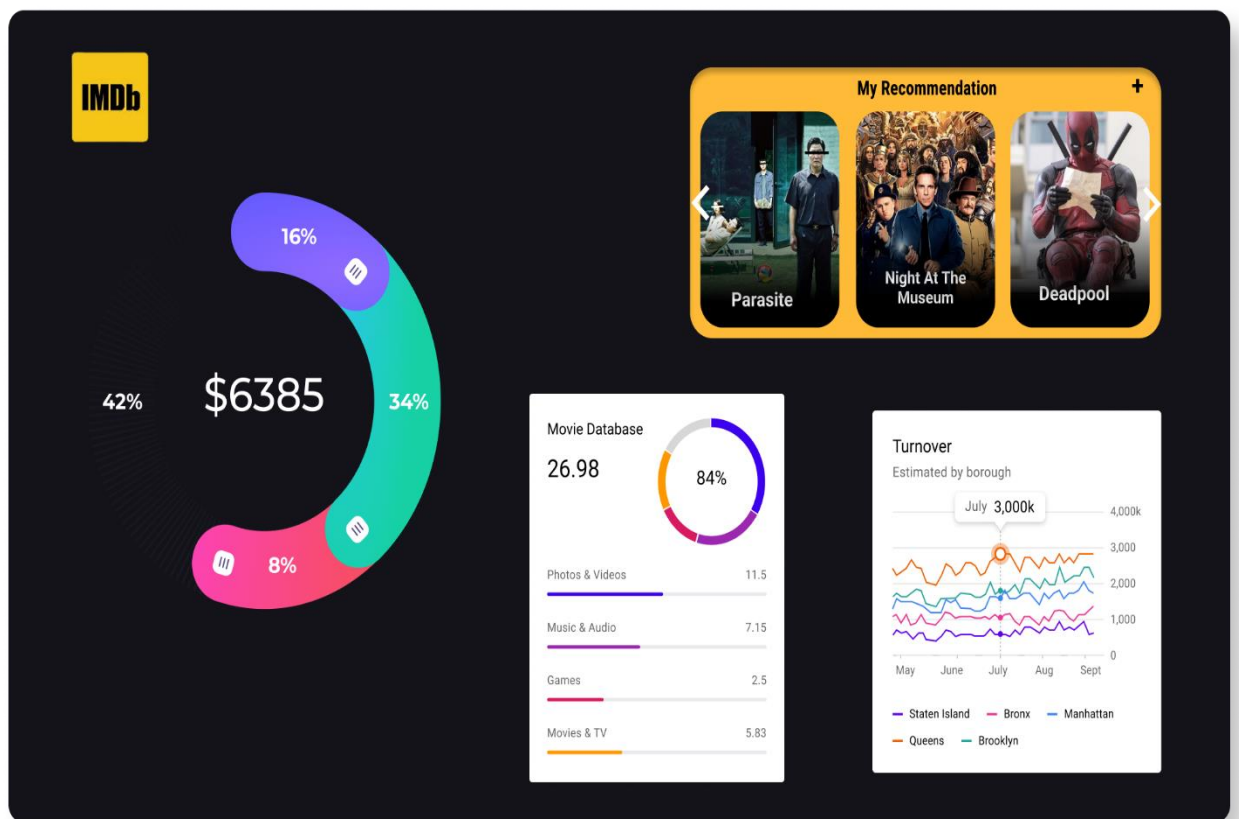


Figure 1 - <https://trainity.link/data/project05>

---

## INDEX

### Contents

|  |    |
|--|----|
| Introduction : .....   | 1  |
| DESCRIPTION .....  | 3  |
| Requirements – .....   | 5  |
| Data Cleaning Task : .....   | 7  |
| Data Analytics Tasks : .....   | 8  |
| A) Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the<br>IMDB score. Q1 - Excel_Answer_File.....     | 8  |
| B) Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the<br>IMDB score. Q2 - Excel_Answer_File..... | 10 |
| C) Language Analysis: Analyze Examine the distribution of movies based on their language..   | 12 |
| D) Director Analysis: Influence of directors on movie ratings. Q4_Answer_Excel_File .....  | 14 |
| E) Budget Analysis: Explore the relationship between movie budgets and their financial<br>success. Q5_Answer_Excel_File .....            | 15 |
| Important Links : .....  | 15 |

---

## DESCRIPTION

### **Problem Statement:**

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 pre-processing 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, you'll explore the data to understand the relationships between different variables. You might look at the correlation between movie ratings and other factors like genre, director, budget, etc. You might also want to consider the year of release, the actors involved, and other relevant factors.

---

## Five 'Whys' Approach:

This technique will help you dig deeper into the problem. For instance, if you find that movies with higher budgets tend to have higher ratings, you can ask "Why?" repeatedly to uncover the root cause. Here's an example:

- Q: "Why do movies with higher budgets tend to have higher ratings?"  
A: They can afford better production quality.
- Q: "Why does better production quality lead to higher ratings?"  
A: It enhances the viewer's experience.
- Q: "Why does an enhanced viewer experience lead to higher ratings?"  
A: Viewers are more likely to rate a movie highly if they enjoyed watching it.
- Q: "Why are viewers more likely to rate a movie highly if they enjoyed watching it?"  
A: Positive experiences lead to positive reviews.
- Q: "Why do positive reviews matter?"  
A: They influence other viewers' decisions to watch the movie, increasing its popularity and success.

## Report and Data Story:

After your analysis, you'll create a report that tells a story with your data. This should include your initial problem, your findings, and the insights you've gained. Use visualizations to help tell your story and make your findings more understandable.

**Remember, as a data analyst, your goal is not just to answer questions but to provide insights that can drive decision-making. Your analysis should aim to provide actionable insights that can help stakeholders make informed decisions.**

---

## Requirements –

### 1) Project Description :

The aim of the project is to find the use the knowledge of Descriptive statistics to find out necessary insights from an excel sheet to grab the details and trends of the IMDB movie dataset which is a place to collect reviews and ratings of various movies and helps people to get the information about the relevance of the same.

After looking at the data we plan a format to operate on the data, tools needed and charts that can be useful for trends etc. while keeping in the questions in mind. The descriptive analysis at various stages of the questions helps us to get the major trends and their effects on ratings given to movies be it based on length, language or directors etc.

### 2) Approach :

I first analyzed the data and looked for null values, blanks, duplicates and treated them using basic functions like delete cells, find & replace, remove blank rows etc. For certain cell values I changed them to suit better with other values like – (for genre separation part) etc. After confirming that the data has little to less outliers (single rating or single genre in movie) and saved the raw data to work on with the operations. Also I removed the most irrelevant columns from the dataset

The analysis based on excel functions, Descriptive analysis, pivot tables has insights at the bottom of the screengrabs to let the others understand the aim of each analysis.

---

### 3) Tech-Stack Used :

Excel – The basic data manipulation, handling and overall pivot charts and the statistics has been handled using MS Excel.

Google Sheets – Used to do the basic data manipulation and to get column stats (gif added)

Word – The report is written in word/docx format using MS Word and then exported to pdf.

Drive – To upload all the essential files attached in the report for reference & pdf upload.

### 4) Insights:

The summary for each query is given with the screenshot but to summarize the overall thing I came to the conclusion that among the total data set we had **5043 data rows, 28 columns & a total of 139145** cells in the given dataset.

After removing the blanks , duplicates and adjusting the non-relevant columns we made the dataset to **3784 rows, 10 columns & 37850 cells** for our final calculation. For question 1 as the genre column had multiple genre in it we further sorted the data for that particular question to **5043 rows, 18 rows (8 genre columns separated) & 49580 cells after splitting** the genre cells value.

For more data insight on the questions look at the respective questions for screenshot and pivot charts.

---

## 5) Result:

The Project has given me a good idea about the importance and vast variety of excel usage which helps us to look deep into plain numbers and generate a visually insightful data which can help business to gain knowledge and prepare for future as well as give out trends to focus on from the numerical data. The statistics section has helped me learn about the various concept which are useful for majority of the operation for handling and displaying basic charts and generate a meaningful dashboard as well as the use of Descriptive stats to get more deeper insights in the data.

## Data Cleaning Task :

Will not explain much just have added the links of gif/process.(These were the basic steps to clean and adjust the data)

[Data Count & Column Stats](#)

[Data Cleaning \(Blanks & Duplicate Removal\)](#)

[Data Cleaning \( Unwanted Column Removal\)](#)

[Data Manipulation \(Column Interchange - to make more readable\)](#)

[Genre Column Separation \(For Q1 - Analysis\)](#)

[Video](#) PPT (Drive Folder)

[Zip](#) (As google drive converts excel sheets to docs and then the connections & plots are lost)

# Data Analytics Tasks :

**A) Movie Genre Analysis:** Analyze the distribution of movie genres and their impact on the IMDB score. [Q1 - Excel Answer File](#)

- **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.

|    |  |                     |                   |                   |                    |                      |
|----|--|---------------------|-------------------|-------------------|--------------------|----------------------|
| 1  | Adding multiple Genre to check results |                     |                   |                   |                    |                      |
| 2  | genres2                                | (All)               |                   |                   |                    |                      |
| 3  | genres3                                | (All)               |                   |                   |                    |                      |
| 4  | genres42                               | (All)               |                   |                   |                    |                      |
| 5  | genres5                                | (All)               |                   |                   |                    |                      |
| 6  | genres6                                | (All)               |                   |                   |                    |                      |
| 7  | genres7                                | (All)               |                   |                   |                    |                      |
| 8  | genres8                                | (All)               |                   |                   |                    |                      |
| 9  | Primary Genre                          |                     | Range of score    |                   |                    |                      |
| 10 | Row Labels                             | Count of imdb_score | Min of imdb_score | Max of imdb_score | Mean of imdb_score | StdDev of imdb_score |
| 11 | Action                                 | 935                 | 2.1               | 9                 | 6.285989305        | 1.078186788          |
| 12 | Adventure                              | 367                 | 2.3               | 8.6               | 6.561307902        | 1.264345826          |
| 13 | Animation                              | 46                  | 4.5               | 8                 | 6.763043478        | 0.945937198          |
| 14 | Biography                              | 206                 | 4.5               | 8.9               | 7.151941748        | 0.489337675          |
| 15 | Comedy                                 | 1026                | 1.9               | 8.8               | 6.164424951        | 1.074567328          |
| 16 | Crime                                  | 252                 | 3.3               | 9.3               | 6.945238095        | 0.75475811           |
| 17 | Documentary                            | 42                  | 1.6               | 8.4               | 6.914285714        | 1.994425087          |
| 18 | Drama                                  | 675                 | 2.1               | 8.8               | 6.822518519        | 0.821539642          |
| 19 | Family                                 | 3                   | 5.7               | 7.9               | 6.5                | 1.216552506          |
| 20 | Fantasy                                | 35                  | 4.3               | 7.9               | 6.234285714        | 0.799378151          |
| 21 | Horror                                 | 156                 | 2.3               | 8.5               | 5.813461538        | 1.015624069          |
| 22 | Musical                                | 2                   | 6.3               | 7.2               | 6.75               | 0.405                |
| 23 | Mystery                                | 23                  | 3.3               | 8.5               | 6.586956522        | 1.23027668           |
| 24 | Romance                                | 2                   | 6.2               | 7.1               | 6.65               | 0.405                |
| 25 | Sci-Fi                                 | 8                   | 5                 | 8.2               | 6.5875             | 1.064107143          |
| 26 | Thriller                               | 3                   | 4.8               | 6.3               | 5.3                | 0.866025404          |
| 27 | Western                                | 3                   | 4.1               | 8.9               | 6.766666667        | 5.973333333          |
| 28 | Grand Total                            | 3784                | 1.6               | 9.3               | 6.461971459        | 1.117810651          |
| 29 |  |                     |                   |                   |                    |                      |

Figure 1) Genre Descriptive stats using pivot table [HD Image](#)

| Descriptive Statistics (For primary Genre) [Using Data Analysis Function in Excel] |              |              |              |              |              |              |              |              |             |              |             |         |              |             |              |          |            |
|--|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------|--------------|-------------|---------|--------------|-------------|--------------|----------|------------|
| Statistics   | Action       | Adventure    | Animation    | Biography    | Comedy       | Crime        | Documentary  | Drama        | Family      | Fantasy      | Horror      | Musical | Mystery      | Romance     | Sci-Fi       | Thriller | Western    |
| Mean   | 5.876190476  | 6.035294118  | 6.54         | 6.965714286  | 5.370769231  | 7.085365854  | 6.786363636  | 6.105769231  | 6.9         | 6.183333333  | 5.973809524 | 7.2     | 6.686666667  | 6.65        | 6.814285714  | 6.3      | 8.1        |
| Standard Error   | 0.235829465  | 0.223112098  | 0.200582485  | 0.186202332  | 0.241397405  | 0.199498181  | 0.300597922  | 0.223714923  | 1           | 0.20900616   | 0.20337054  | 0       | 0.267771521  | 0.45        | 0.329811325  | 0        | 0.8        |
| Median   | 5.9          | 6.1          | 6.7          | 7            | 5.4          | 7.1          | 7.15         | 6.15         | 6.9         | 6.3          | 5.95        | 7.2     | 6.7          | 6.65        | 6.4          | 6.3      | 8.1        |
| Mode   | #N/A         | #N/A         | #N/A         | #N/A         | #N/A         | #N/A         | #N/A         | #N/A         | #N/A        | #N/A         | #N/A        | #N/A    | #N/A         | #N/A        | #N/A         | #N/A     | #N/A       |
| Standard Deviation   | 1.871838348  | 1.593339076  | 1.002912426  | 1.101587853  | 1.946208098  | 1.277411636  | 1.40992923   | 1.613231255  | 1.414213562 | 1.02391689   | 1.317991732 | #DIV/0! | 1.03707464   | 0.636396103 | 0.872598746  | #DIV/0!  | 1.13137085 |
| Sample Variance  | 3.503778802  | 2.538729412  | 1.005833333  | 1.213495798  | 3.787725962  | 1.631780488  | 1.987900433  | 2.602515083  | 2           | 1.048405797  | 1.737102207 | #DIV/0! | 1.07552381   | 0.405       | 0.761428571  | #DIV/0!  | 1.28       |
| Kurtosis   | -1.125806711 | -0.939945976 | -0.939097448 | -1.032947109 | -1.148817444 | -0.786602835 | 2.34645897   | -0.706332748 | #DIV/0!     | -0.892978657 | -0.97450046 | #DIV/0! | 0.024181082  | #DIV/0!     | -1.079431446 | #DIV/0!  | #DIV/0!    |
| Skewness   | -0.063033373 | -0.204231668 | -0.393939313 | -0.08743408  | -0.038496732 | -0.121222876 | -1.512543695 | -0.221550956 | #DIV/0!     | -0.20882416  | 0.076708978 | #DIV/0! | -0.277952541 | #DIV/0!     | 0.717530624  | #DIV/0!  | #DIV/0!    |
| Range  | 6.6          | 5.8          | 3.3          | 3.9          | 6.8          | 5.1          | 5.7          | 6.7          | 2           | 3.6          | 4.9         | 0       | 3.9          | 0.9         | 2.3          | 0        | 1.6        |
| Minimum  | 2.4          | 2.8          | 4.7          | 5            | 2            | 4.2          | 2.7          | 2.1          | 5.9         | 4.3          | 3.6         | 7.2     | 4.6          | 6.2         | 5.9          | 6.3      | 7.3        |
| Maximum  | 9            | 8.6          | 8            | 8.9          | 8.8          | 9.3          | 8.4          | 8.8          | 7.9         | 7.9          | 8.5         | 7.2     | 8.5          | 7.1         | 8.2          | 6.3      | 8.9        |
| Sum  | 370.2        | 307.8        | 163.5        | 243.8        | 349.1        | 290.5        | 149.3        | 317.5        | 13.8        | 148.4        | 250.9       | 7.2     | 100.3        | 13.3        | 47.7         | 6.3      | 16.2       |
| Count  | 63           | 51           | 25           | 35           | 65           | 41           | 22           | 52           | 2           | 24           | 42          | 1       | 15           | 2           | 7            | 1        | 2          |
| Largest(2)   | 8.9          | 8.5          | 7.9          | 8.7          | 8.6          | 9.2          | 8.3          | 8.7          | 5.9         | 7.6          | 8.2         | 1       | 8.1          | 6.2         | 7.7          | 6.3      | 8.9        |
| Smallest(2)  | 2.7          | 3            | 4.8          | 5.1          | 2.1          | 5.1          | 4.1          | 3.4          | 7.9         | 4.4          | 3.9         | 1       | 4.6          | 6.2         | 5.9          | 6.3      | 7.3        |

Note: Mode is shown as NA in calculation as No imdb score seems to be repeated for the genre wise distribution

Figure 2) Descriptive Stats for Primary Genre (Using Data Analysis Tool in Excel) [HD Image](#)



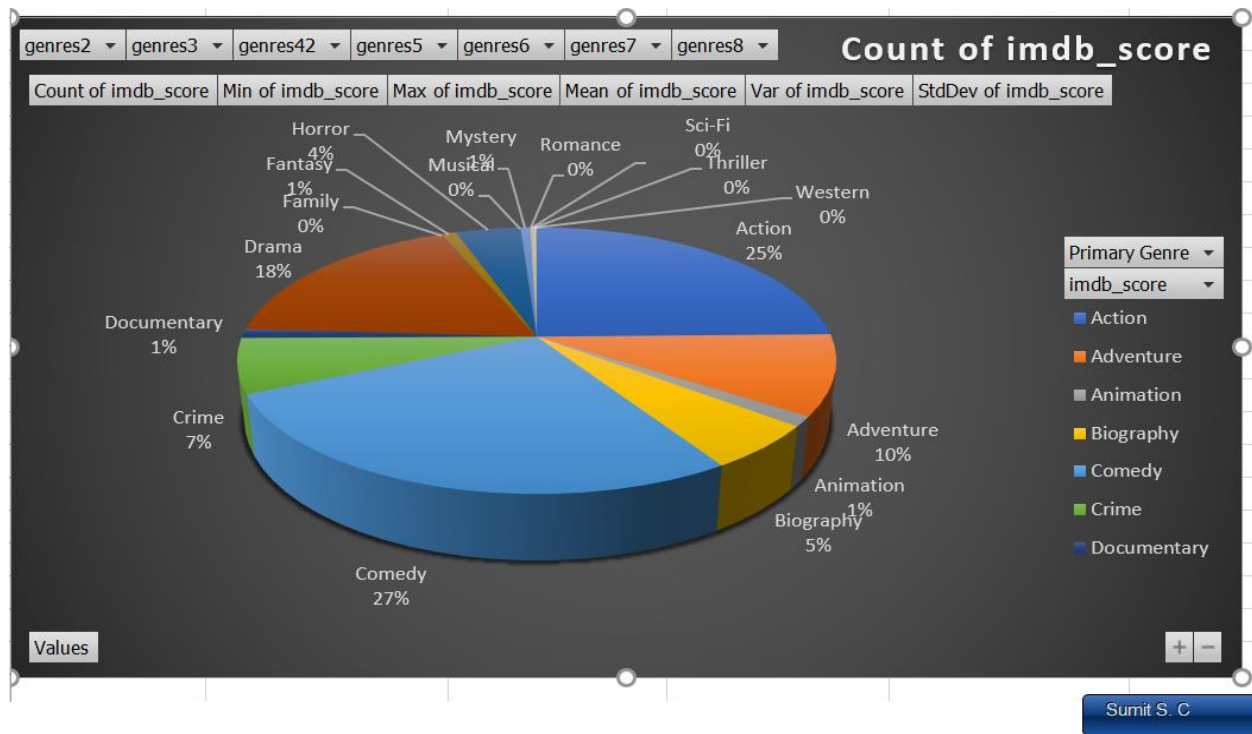


Figure 3) Genre Distribution using Pie – Chart

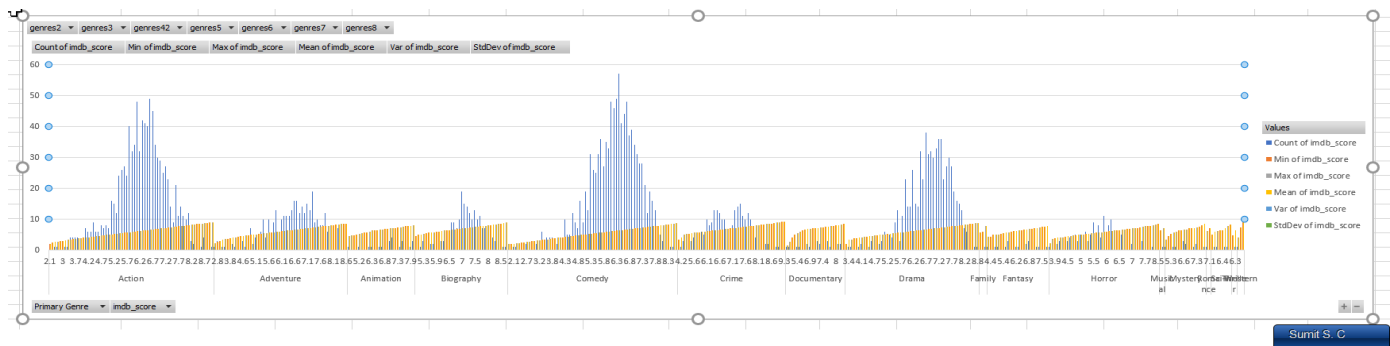


Figure 4) Descriptive Stats BarPlot(Opened Image + ) [HD Image](#)

### Insights:

1. The Genre distribution shows us that **Comedy** is the most popular genre in the dataset with approximately **27% (1026 movies)** and **Action** genre comes at second place with around **25% (935 movies)**. While **Romance & Musical** are both the lowest rated genre in the dataset (**2 movies each**). Note – that the insights are based on the primary genre and so the results would change accordingly as we remove the other genres in calculations.
2. The Analysis shows us that most people love to watch comedy or action movies and they have the most ratings among the genres in the IMDB dataset.

**B) Movie Duration Analysis:** Analyze the distribution of movie durations and its impact on the IMDB score. [Q2 - Excel Answer File](#)

- **Task:** Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

|    | A  | B                   | C                  | D                 | E                 | F                 | G                    | H | I  | J           |
|----|--|---------------------|--------------------|-------------------|-------------------|-------------------|----------------------|---|--|-------------|
| 1  | Creating a Frequency of 25 minutes to group the movie duration |                     |                    |                   |                   |                   |                      |   | <a href="#">Descriptive Stats of Above Range</a> |             |
| 2  |  |                     |                    |                   |                   |                   |                      |   |  |             |
| 3  | Row Labels   | Count of imdb_score | Mean of imdb_score | Min of imdb_score | Max of imdb_score | Var of imdb_score | StdDev of imdb_score |   | Column1  | Column2     |
| 4  | 34-58  | 6                   | 7.216666667        | 6.5               | 7.8               | 0.241666667       | 0.49159604           |   | Mean   | 343.4545455 |
| 5  | 59-83  | 151                 | 6.070198675        | 2.8               | 8.5               | 1.477839294       | 1.215664137          |   | Standard Error                                   | 197.1361949 |
| 6  | 84-108   | 1991                | 6.152285284        | 1.9               | 8.6               | 1.114556333       | 1.055725501          |   | Median   | 26          |
| 7  | 109-133  | 1221                | 6.734807535        | 1.6               | 8.8               | 0.742254656       | 0.861542022          |   | Mode   | 2           |
| 8  | 134-158  | 299                 | 7.184280936        | 3.6               | 9.3               | 0.669450068       | 0.818199284          |   | Standard Deviation                               | 653.8267911 |
| 9  | 159-183  | 67                  | 7.47761194         | 4.8               | 9.2               | 0.862976029       | 0.928965031          |   | Sample Variance                                  | 427489.4727 |
| 10 | 184-208  | 26                  | 7.553846154        | 5.5               | 8.9               | 0.696184615       | 0.834376783          |   | Kurtosis   | 3.903507237 |
| 11 | 209-233  | 12                  | 7.533333333        | 5.8               | 9                 | 0.791515152       | 0.889671373          |   | Skewness   | 2.139598205 |
| 12 | 234-258  | 4                   | 7.9                | 7                 | 8.4               | 0.386666667       | 0.62182527           |   | Range  | 1989        |
| 13 | 259-283  | 2                   | 7                  | 6.3               | 7.7               | 0.98              | 0.989949494          |   | Minimum  | 2           |
| 14 | 284-308  | 3                   | 7.833333333        | 6.6               | 8.5               | 1.143333333       | 1.069267662          |   | Maximum  | 1991        |
| 15 | 309-333  | 2                   | 7.4                | 6.8               | 8                 | 0.72              | 0.848528137          |   | Sum  | 3778        |
| 16 | Grand Total  | 3784                | 6.461971459        | 1.6               | 9.3               | 1.117810651       | 1.057265648          |   | Count  | 11          |
| 17 |  |                     |                    |                   |                   |                   |                      |   | Largest(1)                                       | 1991        |
| 18 |  |                     |                    |                   |                   |                   |                      |   | Smallest(1)                                      | 2           |

Figure 5) Impact of Movie Length(Duration) on scores [HD Image](#)

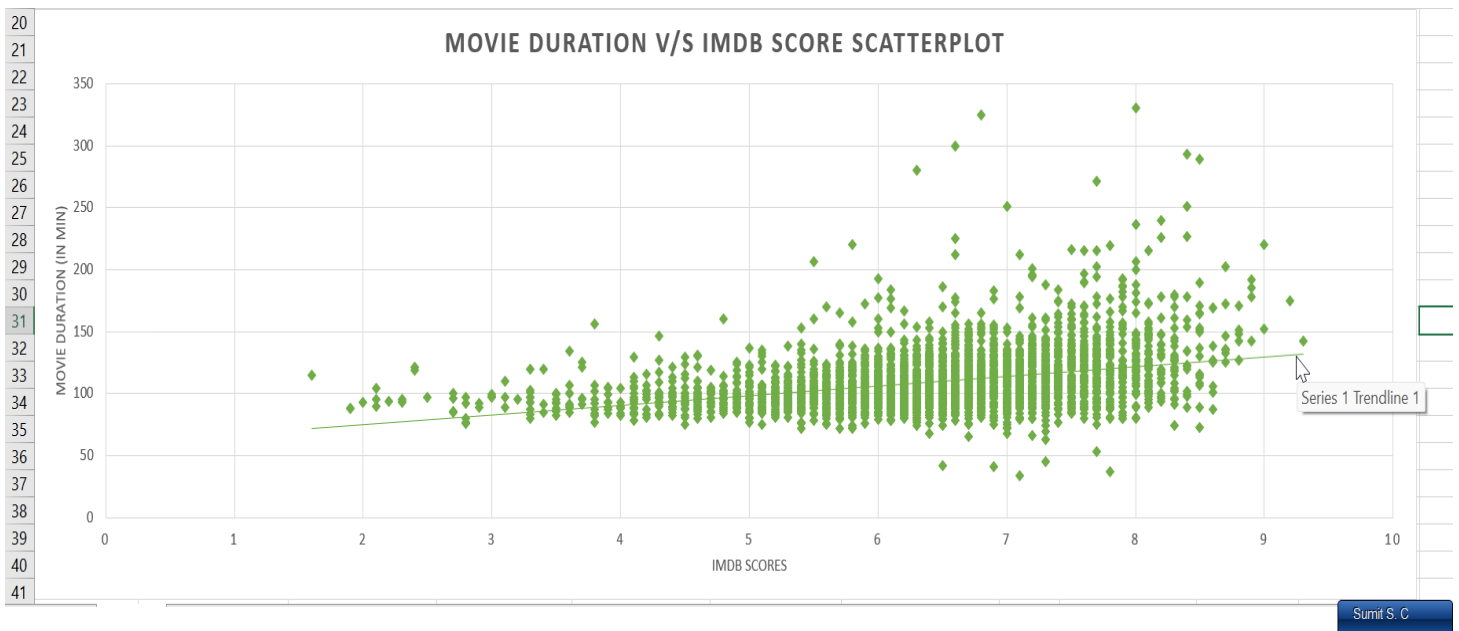


Figure 6) Scatterplot For Length of movie and its impact on scores with a trendline to see the impact [HD Image](#)

---

***Insights:***

1. I have made a frequency table of the duration into 25minutes slots to make the data more readable and get the distribution on plot.
2. From the Pivot table it is evident that most movies are in the range of 84-108minutes long and also in range of 109-133 minutes on average count where people like to rate them on imdb.
3. The analysis shows that the movies in range of 84-133 minutes are among the most rated movies or we could assume are favoured by people more to watch & rate.
4. The analysis can simply be broken down more by just changing the frequency value in group which I have kept to 25 min to get more deep insights on the score v/s length distribution.

## C) Language Analysis: Analyze Examine the distribution of movies based on their language.

### [Q3 Excel Answer File](#)

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

|    | A           | B                   | C                  | D                 | E                 | F                 | G                    | H | I   | J             |
|----|-------------|---------------------|--------------------|-------------------|-------------------|-------------------|----------------------|---|---|---------------|
| 1  | Row Labels  | Count of imdb_score | Mean of imdb_score | Min of imdb_score | Max of imdb_score | Var of imdb_score | StdDev of imdb_score |   |   |               |
| 2  | Aboriginal  | 2                   | 6.95               | 6.4               | 7.5               | 0.605             | 0.777817459          |   | Note : To Get stats of individual language use data analysis on each language<br>Descriptive Stats of Overall |               |
| 3  | Arabic      | 1                   | 7.2                | 7.2               | 7.2               | #DIV/0!           | #DIV/0!              |   |   |               |
| 4  | Aramaic     | 1                   | 7.1                | 7.1               | 7.1               | #DIV/0!           | #DIV/0!              |   |   |               |
| 5  | Bosnian     | 1                   | 4.3                | 4.3               | 4.3               | #DIV/0!           | #DIV/0!              |   | Stats   | Overall Value |
| 6  | Cantonese   | 8                   | 7.2375             | 6.5               | 7.8               | 0.194107143       | 0.440575922          |   | Mean  | 105.0555556   |
| 7  | Czech       | 1                   | 7.4                | 7.4               | 7.4               | #DIV/0!           | #DIV/0!              |   | Standard Error  | 100.0063803   |
| 8  | Danish      | 3                   | 7.9                | 7.3               | 8.3               | 0.28              | 0.529150262          |   | Median  | 2             |
| 9  | Dari        | 2                   | 7.5                | 7.4               | 7.6               | 0.02              | 0.141421356          |   | Mode  | 1             |
| 10 | Dutch       | 3                   | 7.56666667         | 7.1               | 7.8               | 0.163333333       | 0.404145188          |   | Standard Deviation  | 600.0382818   |
| 11 | Dzongkha    | 1                   | 7.5                | 7.5               | 7.5               | #DIV/0!           | #DIV/0!              |   | Sample Variance   | 360045.9397   |
| 12 | English     | 3605                | 6.42147018         | 1.6               | 9.3               | 1.108057216       | 1.052642967          |   | Kurtosis  | 35.98751248   |
| 13 | Filipino    | 1                   | 6.7                | 6.7               | 6.7               | #DIV/0!           | #DIV/0!              |   | Skewness  | 5.998487604   |
| 14 | French      | 37                  | 7.286486486        | 5.8               | 8.4               | 0.31509009        | 0.561328861          |   | Range   | 3604          |
| 15 | German      | 13                  | 7.692307692        | 6.1               | 8.5               | 0.410769231       | 0.640913811          |   | Minimum   | 1             |
| 16 | Hebrew      | 3                   | 7.5                | 7.2               | 8                 | 0.19              | 0.435881894          |   | Maximum   | 3605          |
| 17 | Hindi       | 10                  | 6.76               | 4.8               | 8                 | 1.236             | 1.111755369          |   | Sum   | 3782          |
| 18 | Hungarian   | 1                   | 7.1                | 7.1               | 7.1               | #DIV/0!           | #DIV/0!              |   | Count   | 36            |
| 19 | Icelandic   | 1                   | 6.9                | 6.9               | 6.9               | #DIV/0!           | #DIV/0!              |   | Largest(1)  | 3605          |
| 20 | Indonesian  | 2                   | 7.9                | 7.6               | 8.2               | 0.18              | 0.424264069          |   | Smallest(1)   | 1             |
| 21 | Italian     | 7                   | 7.185714286        | 5.3               | 8.9               | 1.334761905       | 1.155318962          |   |   |               |
| 22 | Japanese    | 12                  | 7.625              | 6                 | 8.7               | 0.809318182       | 0.899621132          |   |   |               |
| 23 | Kazakh      | 1                   | 6                  | 6                 | 6                 | #DIV/0!           | #DIV/0!              |   |   |               |
| 24 | Korean      | 5                   | 7.7                | 7                 | 8.4               | 0.325             | 0.570087713          |   |   |               |
| 25 | Mandarin    | 14                  | 7.021428571        | 5.6               | 7.9               | 0.586428571       | 0.765786244          |   |   |               |
| 26 | Maya        | 1                   | 7.8                | 7.8               | 7.8               | #DIV/0!           | #DIV/0!              |   |   |               |
| 27 | Mongolian   | 1                   | 7.3                | 7.3               | 7.3               | #DIV/0!           | #DIV/0!              |   |   |               |
| 28 | Norwegian   | 4                   | 7.15               | 6.4               | 7.6               | 0.33              | 0.574456265          |   |   |               |
| 29 | Persian     | 3                   | 8.133333333        | 7.5               | 8.5               | 0.303333333       | 0.550757055          |   |   |               |
| 30 | Portuguese  | 5                   | 7.76               | 6.1               | 8.7               | 0.958             | 0.978774744          |   |   |               |
| 31 | Romanian    | 1                   | 7.9                | 7.9               | 7.9               | #DIV/0!           | #DIV/0!              |   |   |               |
| 32 | Russian     | 1                   | 6.5                | 6.5               | 6.5               | #DIV/0!           | #DIV/0!              |   |   |               |
| 33 | Spanish     | 26                  | 7.05               | 5.2               | 8.2               | 0.6826            | 0.826196103          |   |   |               |
| 34 | Swedish     | 1                   | 7.6                | 7.6               | 7.6               | #DIV/0!           | #DIV/0!              |   |   |               |
| 35 | Telugu      | 1                   | 8.4                | 8.4               | 8.4               | #DIV/0!           | #DIV/0!              |   |   |               |
| 36 | Thai        | 3                   | 6.633333333        | 6.2               | 7.1               | 0.203333333       | 0.450924975          |   |   |               |
| 37 | Vietnamese  | 1                   | 7.4                | 7.4               | 7.4               | #DIV/0!           | #DIV/0!              |   |   |               |
| 38 | Zulu        | 1                   | 7.3                | 7.3               | 7.3               | #DIV/0!           | #DIV/0!              |   |   |               |
| 39 | Grand Total | 3784                | 6.461971459        | 1.6               | 9.3               | 1.117810651       | 1.057265648          |   |   |               |

Sumit S. C

Figure 7) Language wise Imdb Score & Stats [HD Image](#)

Note – For Stats of individual movie one can simply use the data analysis function in excel to select the data from language row and then perform the analysis – the one in screenshot is the overall of all languages. Some are missing or shows random things because they cant be determined with the overall data as some places might have 0 score too.

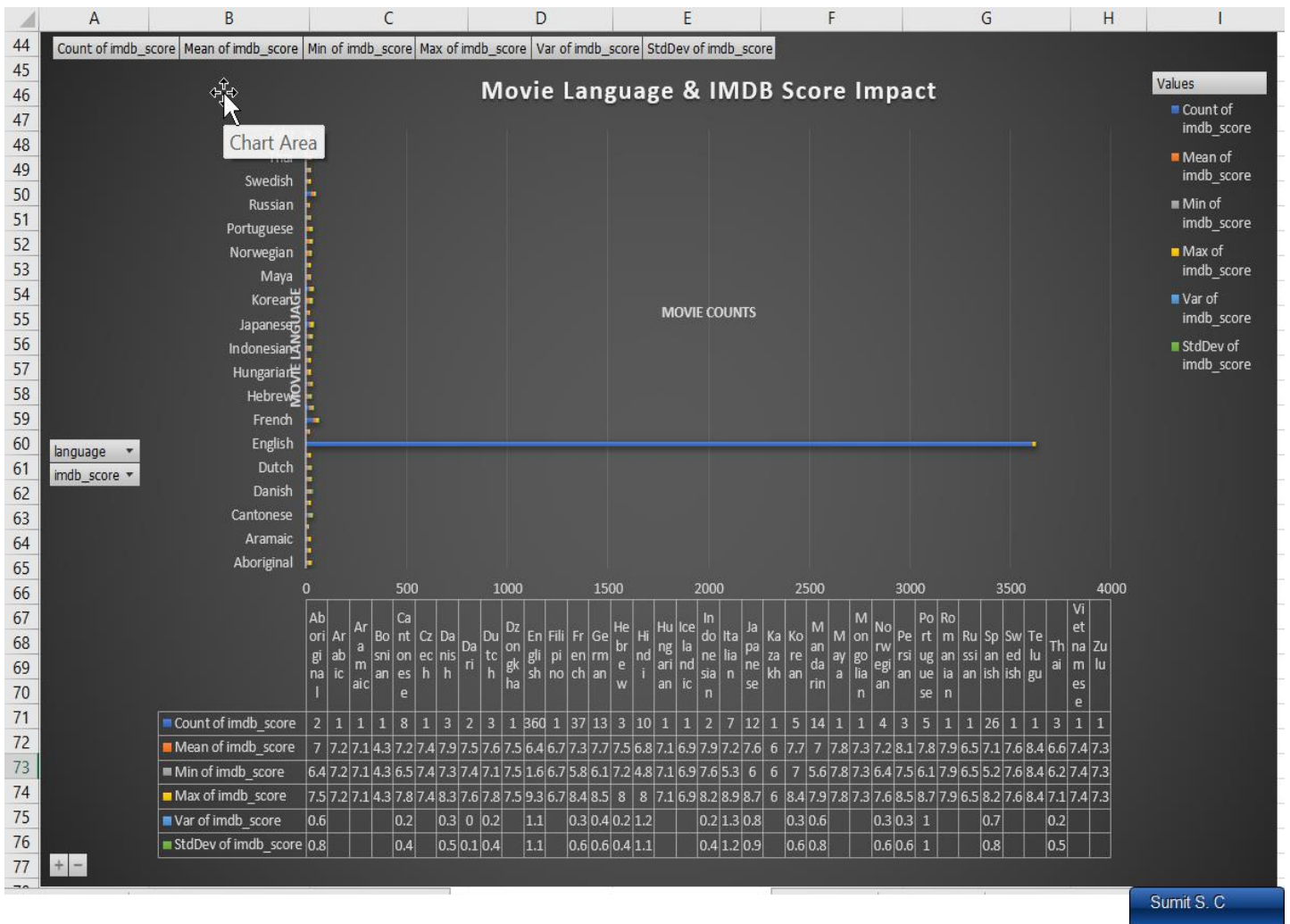


Figure 8) Distribution On plot [HD Image](#)

### Insights:

1. From the analysis of the data it is evident that most of the rated movies on IMDB are from the **English origin (3605 Movie titles)**.
2. Other Prominent languages include **French (37), Spanish (26), Mandarin (14), German (13), Japanese (12)** among other languages.
3. The analysis shows that majority of the user watch **Hollywood movies AKA English (US)** and they are among the most watched languages.

## D) Director Analysis: Influence of directors on movie ratings. [Q4 Answer Excel File](#)

- *Task:* Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

|    | A  | B                          | C                            | D                        | E                             | F                                |
|----|--|----------------------------|------------------------------|--------------------------|-------------------------------|----------------------------------|
| 1  | <b>Movie Director &amp; its Impact On IMDB Score</b> |                            |                              |                          |                               |                                  |
| 2  | <b>Row Labels</b>                                    | <b>Count of imdb_score</b> | <b>Average of imdb_score</b> | <b>Sum of imdb_score</b> | <b>Percentile (WholeData)</b> | <b>Ranking Acc to IMDB_Score</b> |
| 3  | Steven Spielberg                                     | 25                         | 7.544                        | 188.6                    | 0.77%                         | 1                                |
| 4  | Woody Allen  | 19                         | 7                            | 133                      | 0.54%                         | 3                                |
| 5  | Clint Eastwood                                       | 19                         | 7.205263158                  | 136.9                    | 0.56%                         | 2                                |
| 6  | Ridley Scott   | 16                         | 7.13125                      | 114.1                    | 0.47%                         | 5                                |
| 7  | Martin Scorsese                                      | 16                         | 7.675                        | 122.8                    | 0.50%                         | 4                                |
| 8  | Spike Lee  | 15                         | 6.733333333                  | 101                      |                               | 6                                |
| 9  | Steven Soderbergh                                    | 15                         | 6.68                         | 100.2                    |                               | 7                                |
| 10 | Renny Harlin   | 15                         | 5.746666667                  | 86.2                     |                               | 12                               |
| 11 | Tim Burton   | 14                         | 7.05                         | 98.7                     |                               | 8                                |
| 12 | Robert Zemeckis                                      | 13                         | 7.307692308                  | 95                       |                               | 9                                |
| 13 | Oliver Stone   | 13                         | 6.907692308                  | 89.8                     |                               | 11                               |
| 14 | Ron Howard   | 13                         | 6.930769231                  | 90.1                     |                               | 10                               |
| 15 | Robert Rodriguez                                     | 13                         | 5.692307692                  | 74                       |                               | 20                               |
| 16 | Barry Levinson                                       | 13                         | 6.576923077                  | 85.5                     |                               | 13                               |
| 17 | Tony Scott   | 12                         | 6.791666667                  | 81.5                     |                               | 14                               |
| 18 | Michael Bay  | 12                         | 6.616666667                  | 79.4                     |                               | 16                               |
| 19 | Joel Schumacher                                      | 12                         | 6.341666667                  | 76.1                     |                               | 19                               |
| 20 | Shawn Levy   | 11                         | 6.090909091                  | 67                       |                               | 28                               |
| 21 | Rob Reiner   | 11                         | 7.018181818                  | 77.2                     |                               | 18                               |
| 22 | Richard Linklater                                    | 11                         | 7.327272727                  | 80.6                     |                               | 15                               |
| 23 | Chris Columbus                                       | 11                         | 6.654545455                  | 73.2                     |                               | 21                               |
| 24 | Wes Craven   | 10                         | 5.97                         | 59.7                     |                               | 34                               |
| 25 | Sam Raimi  | 10                         | 6.96                         | 69.6                     |                               | 23                               |
| 26 | Paul W.S. Anderson                                   | 10                         | 5.99                         | 59.9                     |                               | 33                               |

Figure 9) Impact of directors on IMDB Score (Long list) [HD Image](#) See the Excel file for more details.

### Insights:

1. From the analysis we get to know the impact of director on the movies IMDB score. Certain Directors holds up as favorites in people mind when they watch the movies maybe due to their excellent cinematography, etc.
2. Certain Directors like **Steven Spielberg, Clint Eastwood, Woody Allen, Martin Scorsese** are among the highest rated director Movies and also have the most number of movies.
3. The count column shows the number of movies of each director while the average shows the average rating for them and the percentile takes the grand total division of the directors rating.
4. Also, the extra rank is made according to the average rating of directors and table is sorted by the number of movies of each director – Total directors are = **1751**

**E) Budget Analysis:** Explore the relationship between movie budgets and their financial success. [Q5 Answer Excel File](#)

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

|    | A  | B          | C                 | D          | E        | F       | G         | H         | I         | J                  |
|----|--|------------|-------------------|------------|----------|---------|-----------|-----------|-----------|--------------------|
| 1  | movie_title                                    | title_year | director_name     | imdb_score | language | country | budget    | gross     | Profit    | Profit(In Million) |
| 2  | Avatar   | 2009       | James Cameron     | 7.9        | English  | USA     | 237000000 | 760505847 | 523505847 | 523.51M            |
| 3  | Jurassic World                                 | 2015       | Colin Trevorrow   | 7          | English  | USA     | 150000000 | 652177271 | 502177271 | 502.18M            |
| 4  | Titanic  | 1997       | James Cameron     | 7.7        | English  | USA     | 200000000 | 658672302 | 458672302 | 458.67M            |
| 5  | Star Wars: Episode IV - A New Hope             | 1977       | George Lucas      | 8.7        | English  | USA     | 11000000  | 460935665 | 449935665 | 449.94M            |
| 6  | E.T. the Extra-Terrestrial                     | 1982       | Steven Spielberg  | 7.9        | English  | USA     | 10500000  | 434949459 | 424449459 | 424.45M            |
| 7  | The Avengers                                   | 2012       | Joss Whedon       | 8.1        | English  | USA     | 220000000 | 623279547 | 403279547 | 403.28M            |
| 8  | The Lion King                                  | 1994       | Roger Allers      | 8.5        | English  | USA     | 45000000  | 422783777 | 377783777 | 377.78M            |
| 9  | Star Wars: Episode I - The Phantom Menace      | 1999       | George Lucas      | 6.5        | English  | USA     | 115000000 | 474544677 | 359544677 | 359.54M            |
| 10 | The Dark Knight                                | 2008       | Christopher Nolan | 9          | English  | USA     | 185000000 | 533316061 | 348316061 | 348.32M            |
| 11 | The Hunger Games                               | 2012       | Gary Ross         | 7.3        | English  | USA     | 78000000  | 407999255 | 329999255 | 330M               |
| 12 | Deadpool                                       | 2016       | Tim Miller        | 8.1        | English  | USA     | 58000000  | 363024263 | 305024263 | 305.02M            |
| 13 | The Hunger Games: Catching Fire                | 2013       | Francis Lawrence  | 7.6        | English  | USA     | 130000000 | 424645577 | 294645577 | 294.65M            |
| 14 | Jurassic Park                                  | 1993       | Steven Spielberg  | 8.1        | English  | USA     | 63000000  | 356784000 | 293784000 | 293.78M            |
| 15 | Despicable Me 2                                | 2013       | Pierre Coffin     | 7.5        | English  | USA     | 76000000  | 368049635 | 292049635 | 292.05M            |
| 16 | American Sniper                                | 2014       | Clint Eastwood    | 7.3        | English  | USA     | 58800000  | 350123553 | 291323553 | 291.32M            |
| 17 | Finding Nemo                                   | 2003       | Andrew Stanton    | 8.2        | English  | USA     | 94000000  | 380838870 | 286838870 | 286.84M            |
| 18 | Shrek 2  | 2004       | Andrew Adamson    | 7.2        | English  | USA     | 150000000 | 436471036 | 286471036 | 286.47M            |
| 19 | The Lord of the Rings: The Return of the King  | 2003       | Peter Jackson     | 8.9        | English  | USA     | 94000000  | 377019252 | 283019252 | 283.02M            |
| 20 | Star Wars: Episode VI - Return of the Jedi     | 1983       | Richard Marquand  | 8.4        | English  | USA     | 32500000  | 309125409 | 276625409 | 276.63M            |
| 21 | Forrest Gump                                   | 1994       | Robert Zemeckis   | 8.8        | English  | USA     | 55000000  | 329691196 | 274691196 | 274.69M            |
| 22 | Star Wars: Episode V - The Empire Strikes Back | 1980       | Irvin Kershner    | 8.8        | English  | USA     | 18000000  | 290158751 | 272158751 | 272.16M            |
| 23 | Home Alone                                     | 1990       | Chris Columbus    | 7.5        | English  | USA     | 18000000  | 285761243 | 267761243 | 267.76M            |
| 24 | Star Wars: Episode III - Revenge of the Sith   | 2005       | George Lucas      | 7.6        | English  | USA     | 113000000 | 380262555 | 267262555 | 267.26M            |
| 25 | Spider-Man                                     | 2002       | Sam Raimi         | 7.3        | English  | USA     | 139000000 | 403706375 | 264706375 | 264.71M            |
| 26 | Minions  | 2015       | Kyle Balda        | 6.4        | English  | USA     | 74000000  | 336029560 | 262029560 | 262.03M            |

Figure 10) Movie Budget Analysis (Top 25 Movies for eg) [HD Image](#)

### Insights:

1. By subtracting the Gross with Budget we get the actual Profit/Loss of the movies, we sorted it by largest value and then took the top 15 movies to analyse our data.
2. The top movie from our data is **Avatar** at a staggering **\$535Million in profits** it had a **budget of \$237M** so its almost 2X of investment (Note – the data is not upto mark as Avatar is supposed to be around \$2.2B according to sources but our data might be just of US & of old origins).
3. The other top 25 movies are shown in the screenshots.

### Important Links :

---

[Drive Folder Link](#)

[Individual Excel Sheets](#) (Individual Questions File)

[Final Excel sheet](#) (cleaned dataset)

Word File Link & Pdf File (Will Be in the [drive folder](#) – can't add before I upload the file)

[Video Presentation](#)

---

# Thank You

---