TRAINITY

**PROJECT – 5**

**IMDB MOVIE ANALYSIS**

Submitted By

**BOJJA JAHNAVI**

Data Analytics Trainee

Submitted On:- 18-11-2023

**IMDB MOVIE ANALYSIS**

**Project Description:**

IMDB Movie Analysis is the description of the released movies till now. It contains the details of all movies to get the clear information about them. The analysis is done by undergoing through the details of each movie like director, title, duration, budget, artists, ratings, profits, gross earnings etc. It also has the IMDB score to estimate the position of the movie.

As a Data Analyst, we have to start the project by cleaning the given data and analyzing it.

**Data Cleaning:** The given data must be cleaned prior to analyzing and visualizing. So in the information provided in the form of spread sheet, remove the blank cells to avoid mistakes in the solutions while deriving them. Remove duplicate cells, convert the data into useful datatypes, remove the unnecessary columns. After performing all these activites the given data gets cleaned and ready for analyzing and deriving the solutions.

**Data Analysing:** Analyse the information of the movies like budget spent on each movie, their Imdb score, ratings, profits. All these information is needed to know the top rated movies. Correlation of two of these variables is done to eatimate the relation between each of them.

**Approach:**

This project is initiated by downloading the dataset in the form of excel sheet. Perform the project in Microsoft Excel by using various functions. After cleaning the data procede to derive the solutions to the given tasks.

1. **Movie Genre Analysis: Genre analysis is done by identifying various genres movies using countif function. Descriptive analysis of the genres is done by finding the mean, median, mode, variance, standard deviation.**

**Function:**

=COUNTIF(E:E,"\*" & T4 & "\*")

The functions for descriptive analysis are

=Averageif (range, criteria, average\_range)

=median(if(isnumber(search(criteria, range)),median\_range))

=mode(if(isnumber(search(criteria, range)),mode\_range))

=var(if(isnumber(search(criteria, range)),var\_range))

=Stdev(if(isnumber(search(criteria, range)),stdev\_range))

=maxifs(max\_range, range, criteria)

=minifs(min\_range, range, criteria)

**Output:**

The no of **genres are 19** and each genre has many movies.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Genre** | **No of Movies** | **Mean IMDB Score** | **Median IMDB Score** | **Mode IMDB Score** | **Max IMDB Score** | **Min IMDB Score** | **Var of IMDB** |
| Action | 934 | 6.484582 | 6.5 | 6.7 | 9 | 2.2 | 0.988744 |
| Adventure | 764 | 6.515838 | 6.5 | 6.4 | 9 | 2.2 | 0.951557 |
| Animation | 197 | 6.404061 | 6.4 | 6.8 | 8.3 | 3.7 | 0.818147 |
| Biography | 241 | 6.43195 | 6.5 | 6.3 | 8.6 | 2.7 | 0.95185 |
| Comedy | 1489 | 6.385829 | 6.5 | 6.6 | 9.3 | 1.6 | 1.201674 |
| Crime | 702 | 6.423362 | 6.5 | 6.5 | 9.3 | 2.1 | 1.044817 |
| Drama | 1908 | 6.444654 | 6.6 | 6.7 | 9.2 | 1.9 | 1.179232 |
| Family | 441 | 6.402268 | 6.4 | 6.4 | 8.4 | 2.3 | 1.112995 |
| Fantasy | 495 | 6.471919 | 6.5 | 6.6 | 8.9 | 2.3 | 1.028906 |
| Horror | 377 | 6.542971 | 6.6 | 7.2 | 8.9 | 2.4 | 1.044957 |
| History | 152 | 6.393421 | 6.6 | 6.7 | 8.5 | 2.1 | 1.015586 |
| Mystery | 376 | 6.575532 | 6.6 | 7.1 | 9.3 | 2.1 | 1.126546 |
| Musical | 102 | 6.30098 | 6.4 | 6.6 | 8.6 | 2.7 | 1.247623 |
| Music | 247 | 6.453846 | 6.6 | 6.6 | 9 | 2.7 | 1.235016 |
| Romance | 865 | 6.463006 | 6.6 | 5.9 | 9 | 1.6 | 1.34944 |
| Sci-Fi | 482 | 6.559751 | 6.7 | 7.1 | 9 | 2.2 | 1.158543 |
| Thriller | 1085 | 6.460184 | 6.6 | 6.7 | 8.9 | 2.1 | 1.076716 |
| War | 158 | 6.401266 | 6.55 | 6.7 | 8.7 | 2.1 | 1.113119 |
| Western | 58 | 6.543103 | 6.65 | 7.1 | 8.3 | 3.5 | 0.835127 |

These are the descriptive analysis of genres and IMDB Scores.

1. **Movie Duration Analysis: Movie Duration and IMDB score is analyzed by executing descriptive functions like mean, median, standard deviation. Draw a trendline to know the strength of the relation between duration and score.**

**Function:**

**=average(C:C)**

**=median(C:C)**

**=var(C:C)**

**=stdev(C:C)**

**A scatterd plot is drawn by selecting IMDB Scores to analyze relation between duaration and scores.**

**Output:**

|  |  |
| --- | --- |
| **Descriptive Analysis** | **Values** |
| Mean | 109.8146 |
| Median | 105 |
| Standard Deviation | 22.7563 |

The regression value of the trend line is  **= 0.7356**

1. **Language Analysis: As IMDB has the details of all movies find the distinct languages used and count the no of movies each language exists. Analyse the descriptive analysis of the movie by comparing with the languages and imdb scores.**

**Function:**

**=countif(range, criteria)**

**For Descriptive Analysis**

**=average(C:C)**

**=median(C:C)**

**=stdev(C:C)**

**Output:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row Labels** | **No of Movies** | **Mean of IMDB** | **Median of IMDB** | **Standard Deviation of IMDB** |
| Aboriginal | 2 | 6.95 | 6.95 | 0.777817 |
| Arabic | 1 | 7.2 | 7.2 | #DIV/0! |
| Aramaic | 1 | 7.1 | 7.1 | #DIV/0! |
| Bosnian | 1 | 4.3 | 4.3 | #DIV/0! |
| Cantonese | 8 | 7.2375 | 7.3 | 0.440576 |
| Czech | 1 | 7.4 | 7.4 | #DIV/0! |
| Danish | 3 | 7.9 | 8.1 | 0.52915 |
| Dari | 2 | 7.5 | 7.4 | 0.732319 |
| Dutch | 3 | 7.566667 | 7.8 | 0.404145 |
| Dzongkha | 1 | 7.5 | 7.5 | #DIV/0! |
| English | 3596 | 6.421023 | 6.5 | 1.052995 |
| Filipino | 1 | 6.7 | 6.7 | #DIV/0! |
| French | 37 | 7.286486 | 7.2 | 0.561329 |
| German | 13 | 7.692308 | 7.7 | 0.640913 |
| Hebrew | 3 | 7.5 | 7.3 | 0.43589 |
| Hindi | 10 | 6.76 | 7.05 | 1.111755 |
| Hungarian | 1 | 7.1 | 7.1 | #DIV/0! |
| Icelandic | 1 | 6.9 | 6.9 | #DIV/0! |
| Indonesian | 2 | 7.9 | 7.9 | 0.424264 |
| Italian | 7 | 7.185714 | 7 | 1.155319 |
| Japanese | 12 | 7.625 | 7.8 | 0.899621 |

These are some of the languages, no of movies and there descriptive analysis.

1. **Director Analysis: Best directors are found by finding their average imdb score. Percentile of each director is calculated to find the top directors.**

**Function:**

**=averageif(A:A,S113,Q:Q)**

**=percentrank.exc(T113:T1858,T113)**

**Output:**

|  |  |  |
| --- | --- | --- |
| **Director** | **Average\_imdb** | **percentile** |
| James Cameron | 7.914285714 | 0.976 |
| Gore Verbinski | 6.985714286 | 0.722 |
| Sam Mendes | 7.457142857 | 0.891 |
| Christopher Nolan | 8.425 | 0.995 |
| Andrew Stanton | 7.733333333 | 0.953 |
| Sam Raimi | 6.96 | 0.718 |
| Nathan Greno | 7.8 | 0.958 |
| Joss Whedon | 7.866666667 | 0.969 |
| David Yates | 7.2 | 0.812 |

**These are the percentile scores of some of the directors.**

1. Budget Analysis: Budget Analysis is done by relation between the gross earnings and budget of each movie. Profit of the movies is obtained by calculating them and top movies are found. Correlation of two variables like gross and budget is calculated.

**Function:**

Profit= Gross – Budget

=Correlation(variable1,variable2)

**Output:**

Correlation= 0.906064

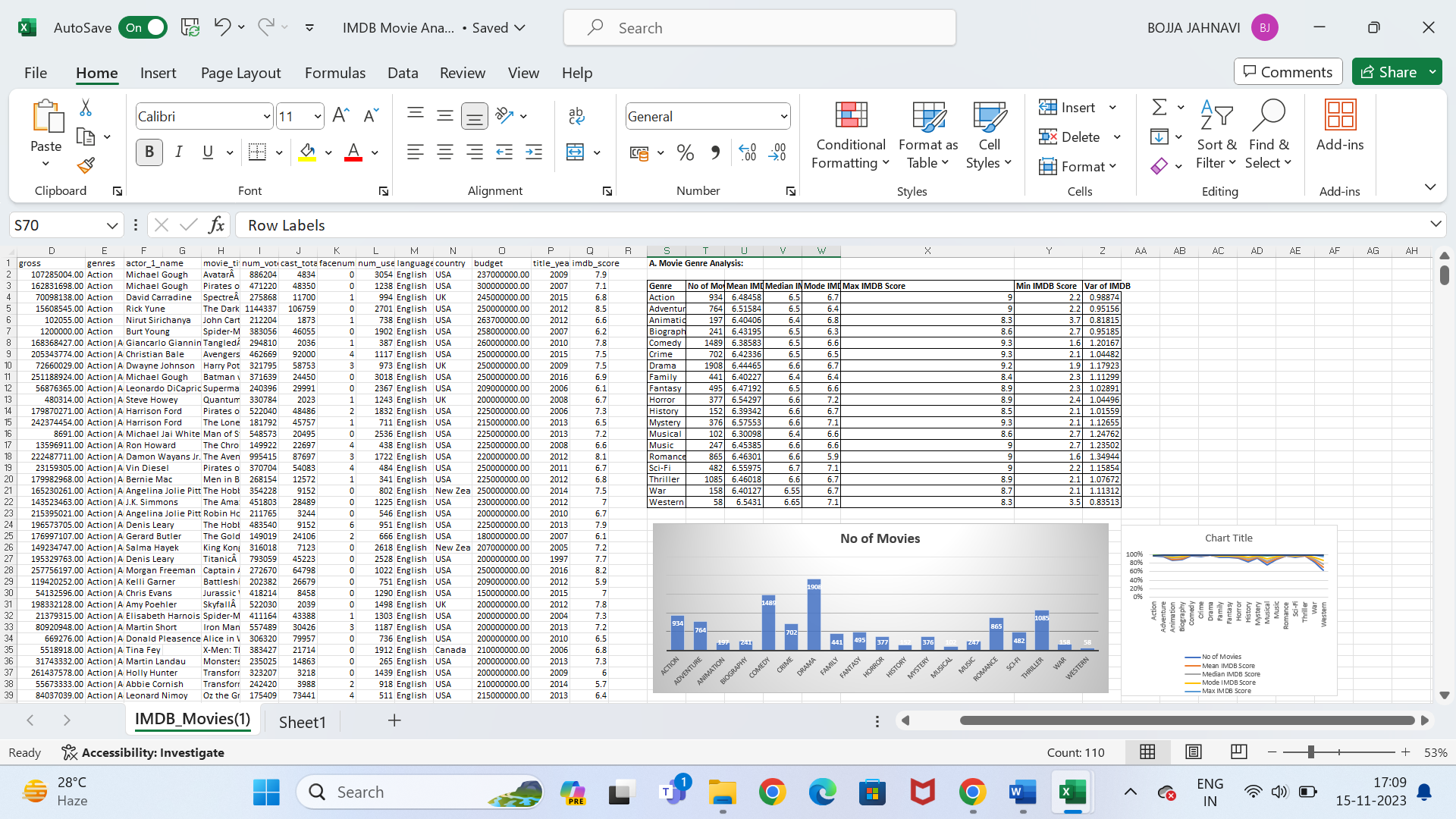
List of Top Movies

|  |  |
| --- | --- |
| **Movies** | **Profits in Millions** |
| **AvatarÂ** | 523505847 |
| **Jurassic WorldÂ** | 502177271 |
| **TitanicÂ** | 458672302 |
| **Star Wars: Episode IV - A New HopeÂ** | 449935665 |
| **E.T. the Extra-TerrestrialÂ** | 424449459 |
| **The Lion KingÂ** | 377783777 |
| **The Jungle BookÂ** | 375290282 |
| **Star Wars: Episode I - The Phantom MenaceÂ** | 359544677 |
| **The Dark KnightÂ** | 348316061 |
| **The Twilight Saga: Breaking Dawn - Part 2Â** | 344597846 |
|  |  |

**Tech-Stack Used:** The Project “IMDB Movie Analysis” is done in Microsoft Excel 2021. This is very simple to learn and handle for beginners. It has every function to solve a question and derive the answer. Therfore it is flexible to use and is free of cost.

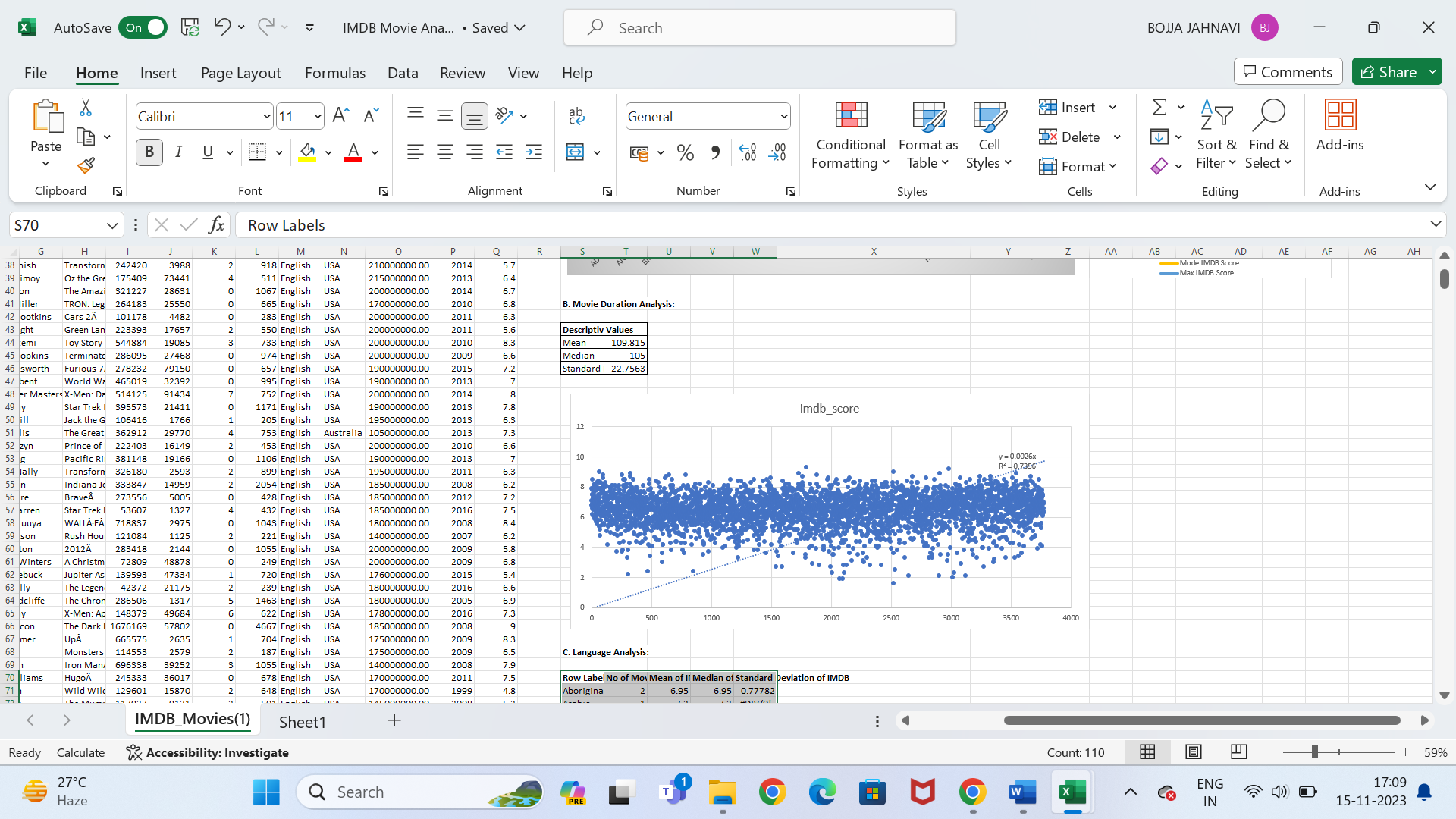
**Insights:**

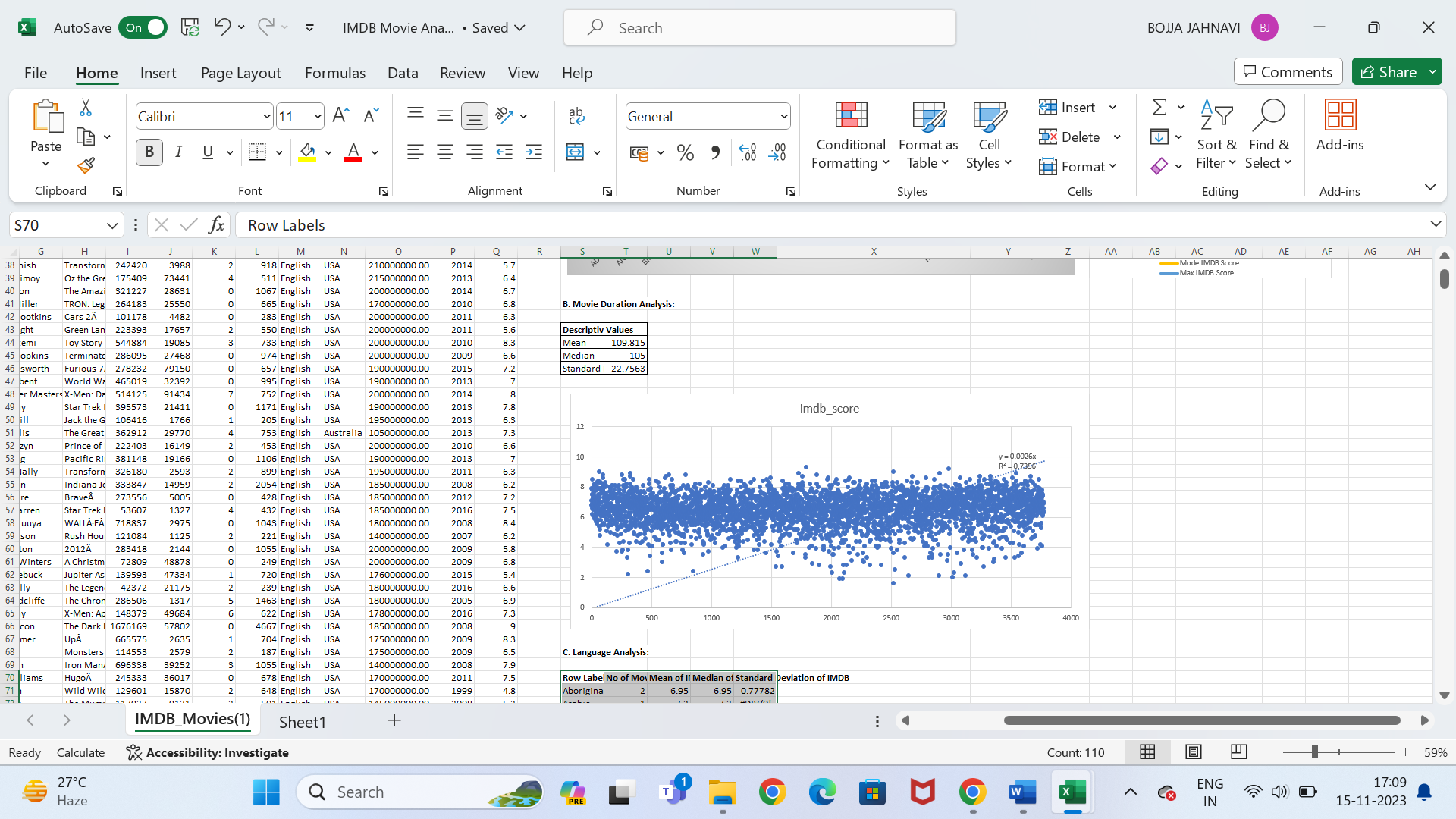
1. **Movie Genre Analysis:**

****

There are total 19 genres. Drama and Comedy Genre has the highest movies.

1. **Movie Duration Analysis:**

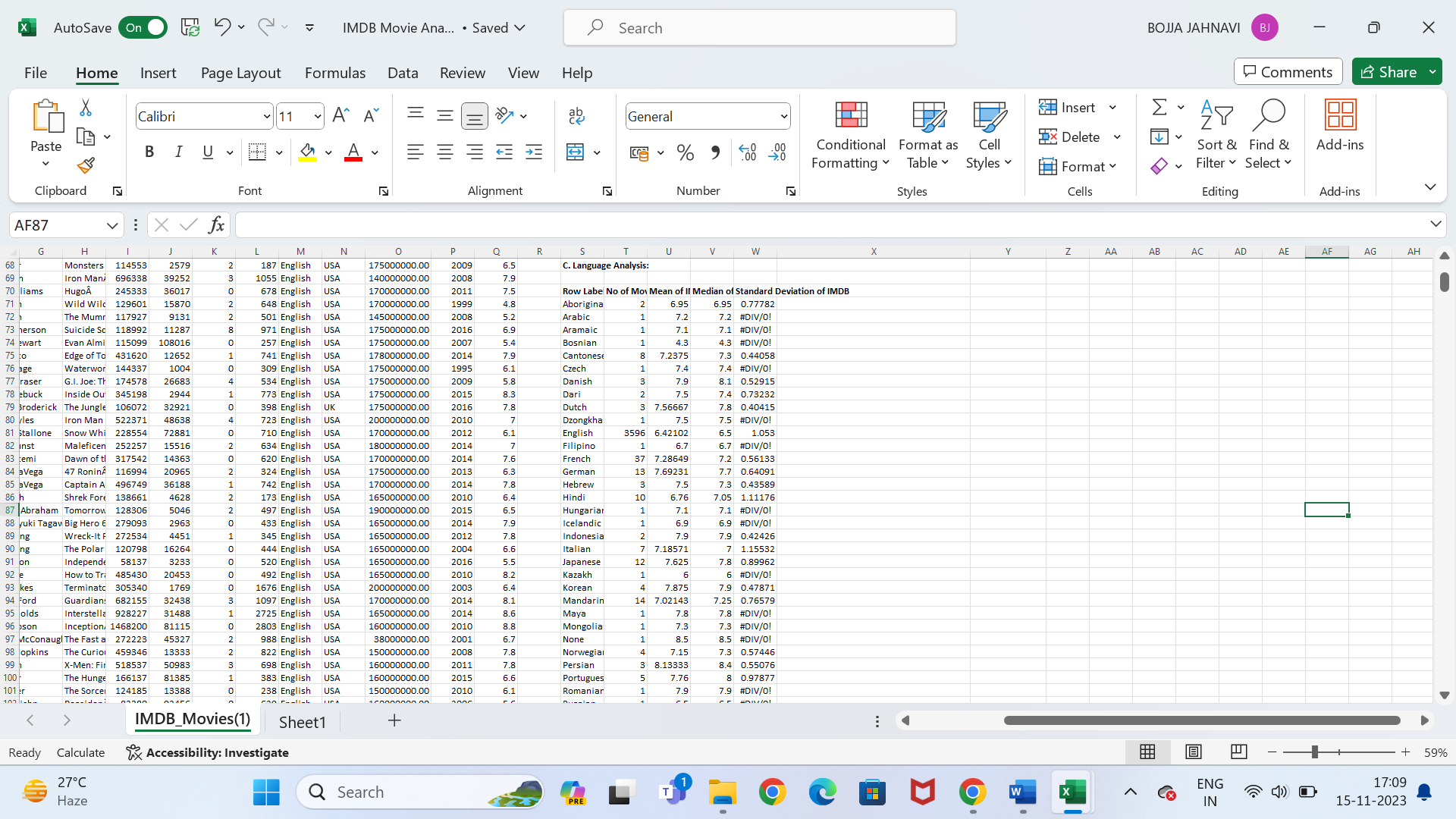
****



The mean, median, mode of duration is **109.815, 105, 22.7563**

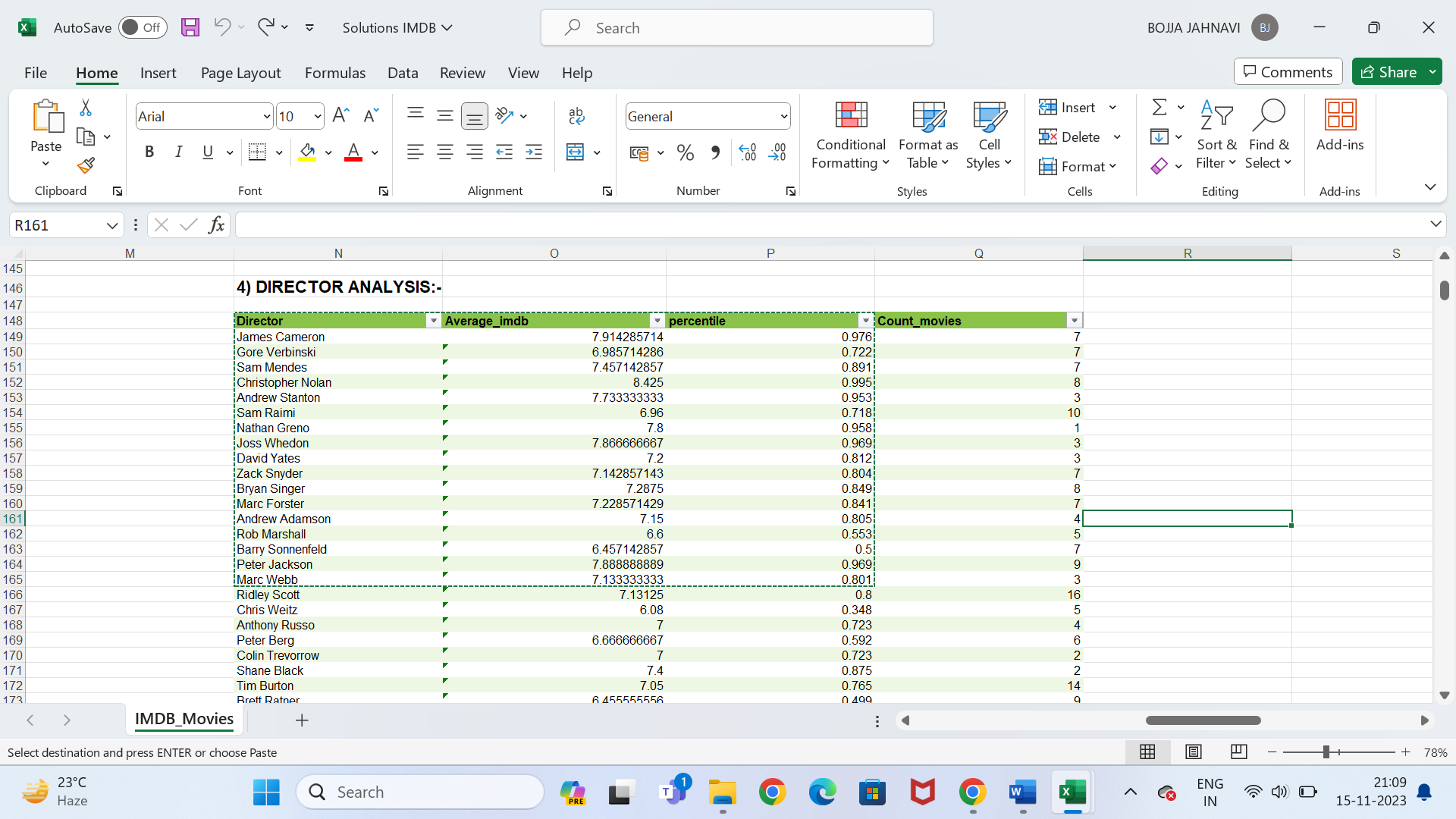
Regression square of the trend line of imdb scores is **0.7356**

1. **Language Analysis:**

****

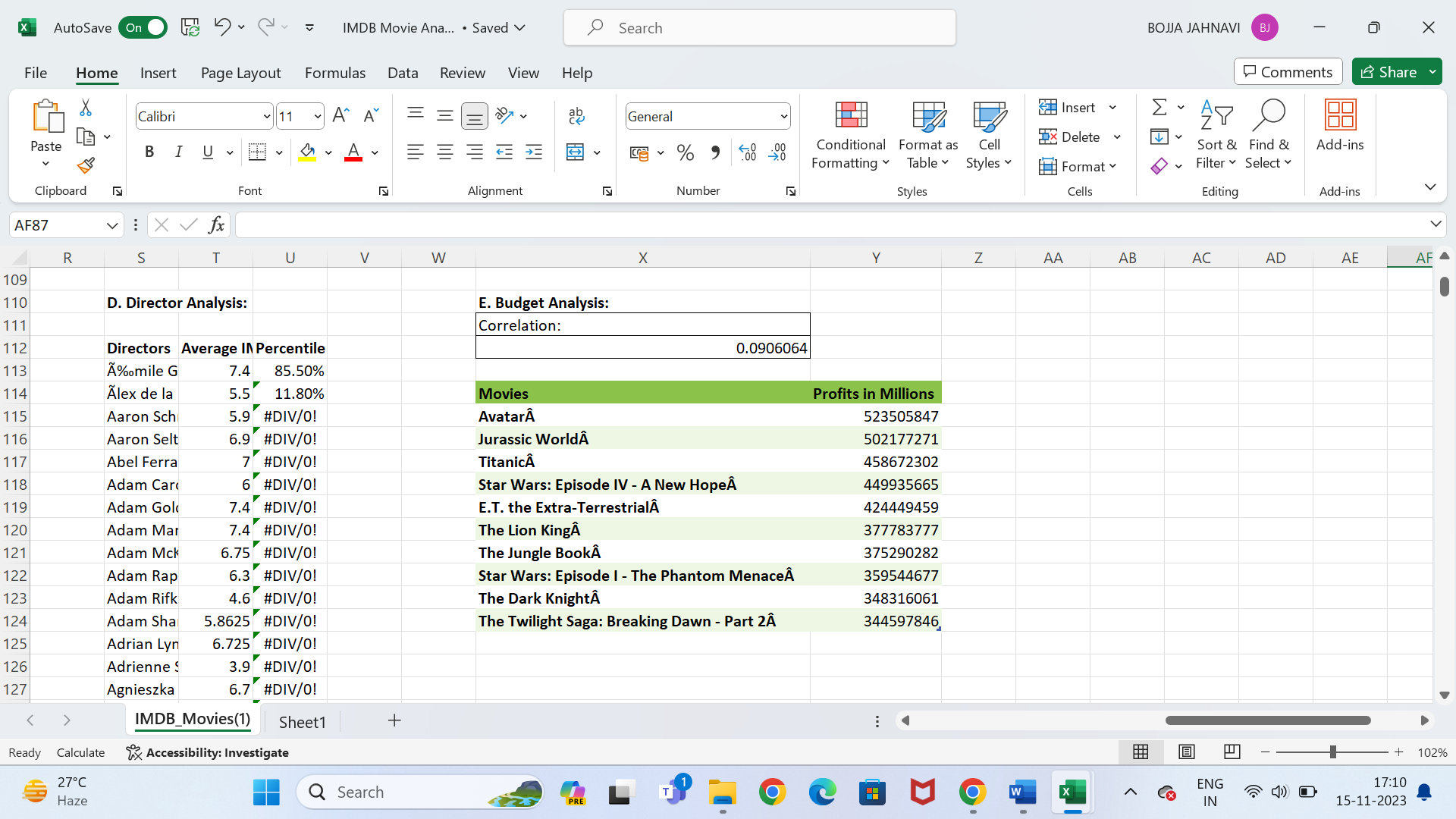
These are the distributions of various languages of the movies.

1. **Director Analysis:**

****

According to the imdb scores the overall top directors are anlysed by obtaining the percentile scores.

1. **Budget Analysis:**



The correlation between Gross and Budget is **0.0906064**

The Highest profited movie is **Avatar.**

**Results:** From the project “IMDB Movie Analysis” I have learnt the advanced excel functions. The overall ratings and top movies are found using the formulas. By this project we can derive meaningful insights from the given data by analysing and coverting into visualization charts. This project improved my excel skills in analytics of large data.

My Excel Worksheet:

<https://docs.google.com/spreadsheets/d/1YkDXB6iSUX6GplfyY3caKAWNN0L0sGpu/edit?usp=sharing&ouid=101880124803050791429&rtpof=true&sd=true>

Video Presentation:

<https://drive.google.com/file/d/1OvfH644Ptc2-bcL3W76mhyeGWgMT6l_b/view?usp=sharing>