**IMDB Movie Analysis**

**Statistics Final Project- 1**

**Project Description**

In this project, we will perform statistical analysis in order to understand IMBD.

**Approach**

Using MS Excel we will do statistical analysis of data records in IMBD Database

**Tech-Stack Used: Microsoft Excel 2016**

* MS Excel 2016 is a spreadsheet program where one can record data in the form of tables.
* It is easy to analyse numerical data in an Excel spreadsheet.
* It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA).

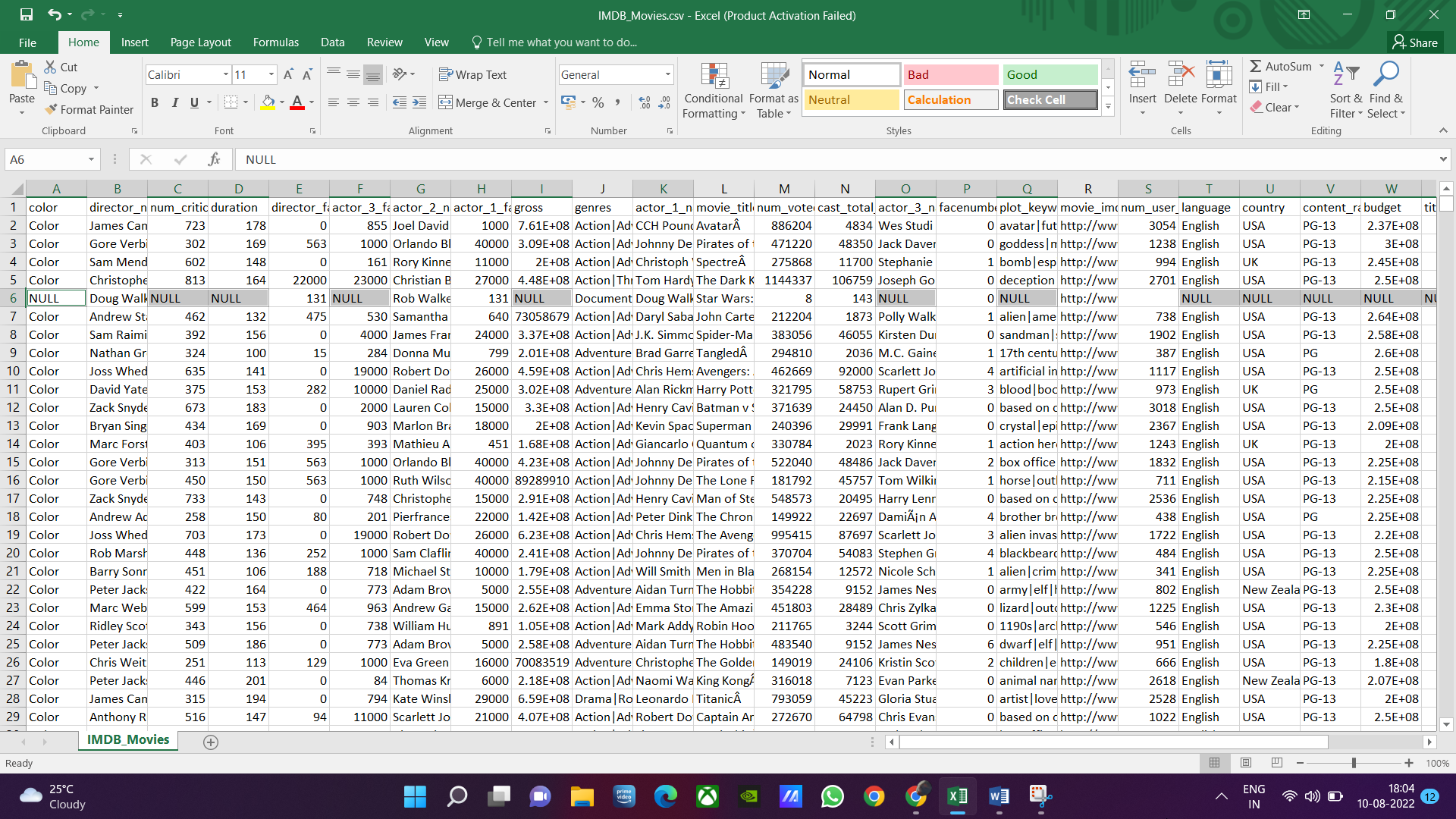
**Insights:**

1. **Cleaning the data:** This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

**Soln:**

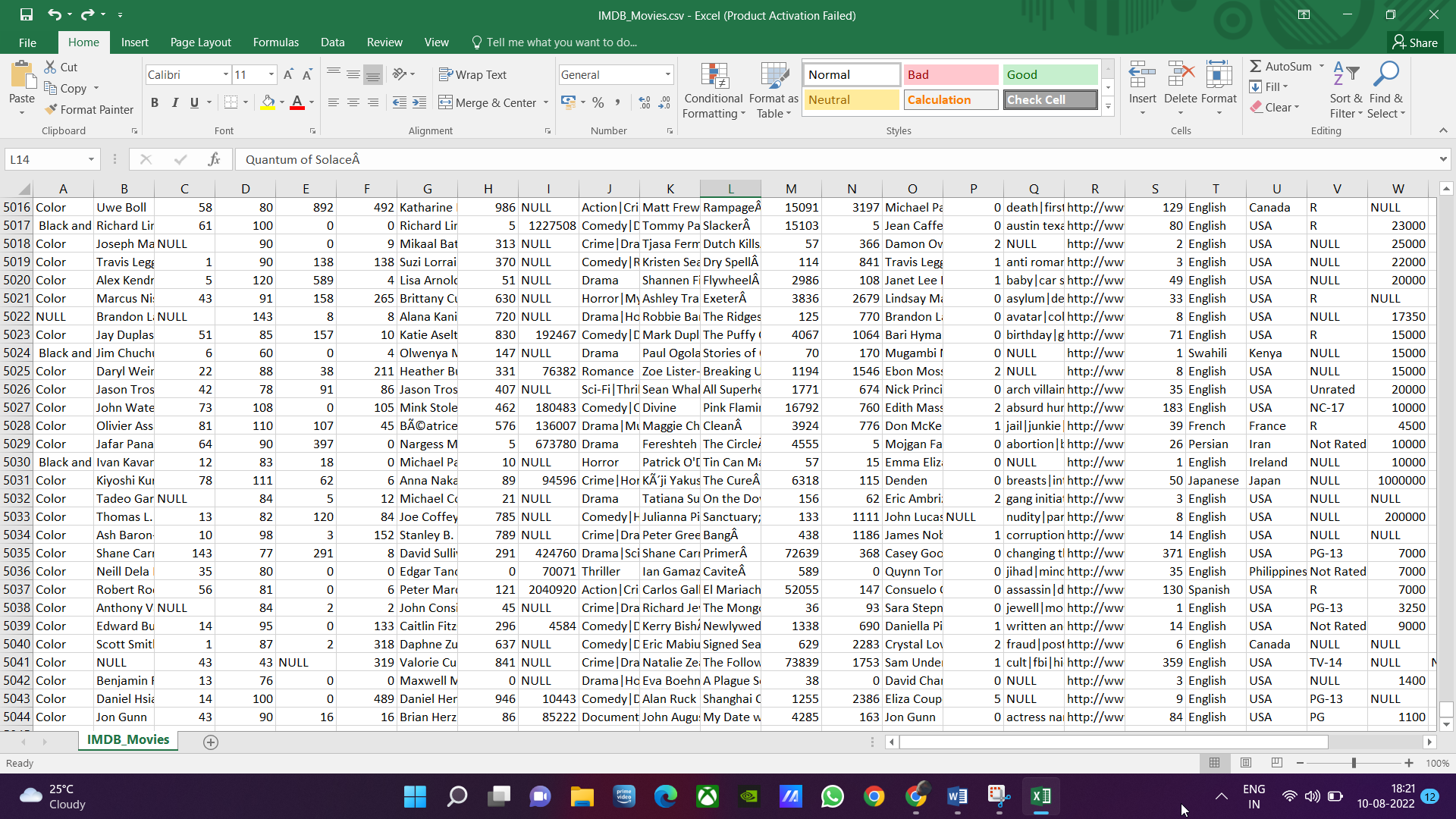
Select and treat all blank cells

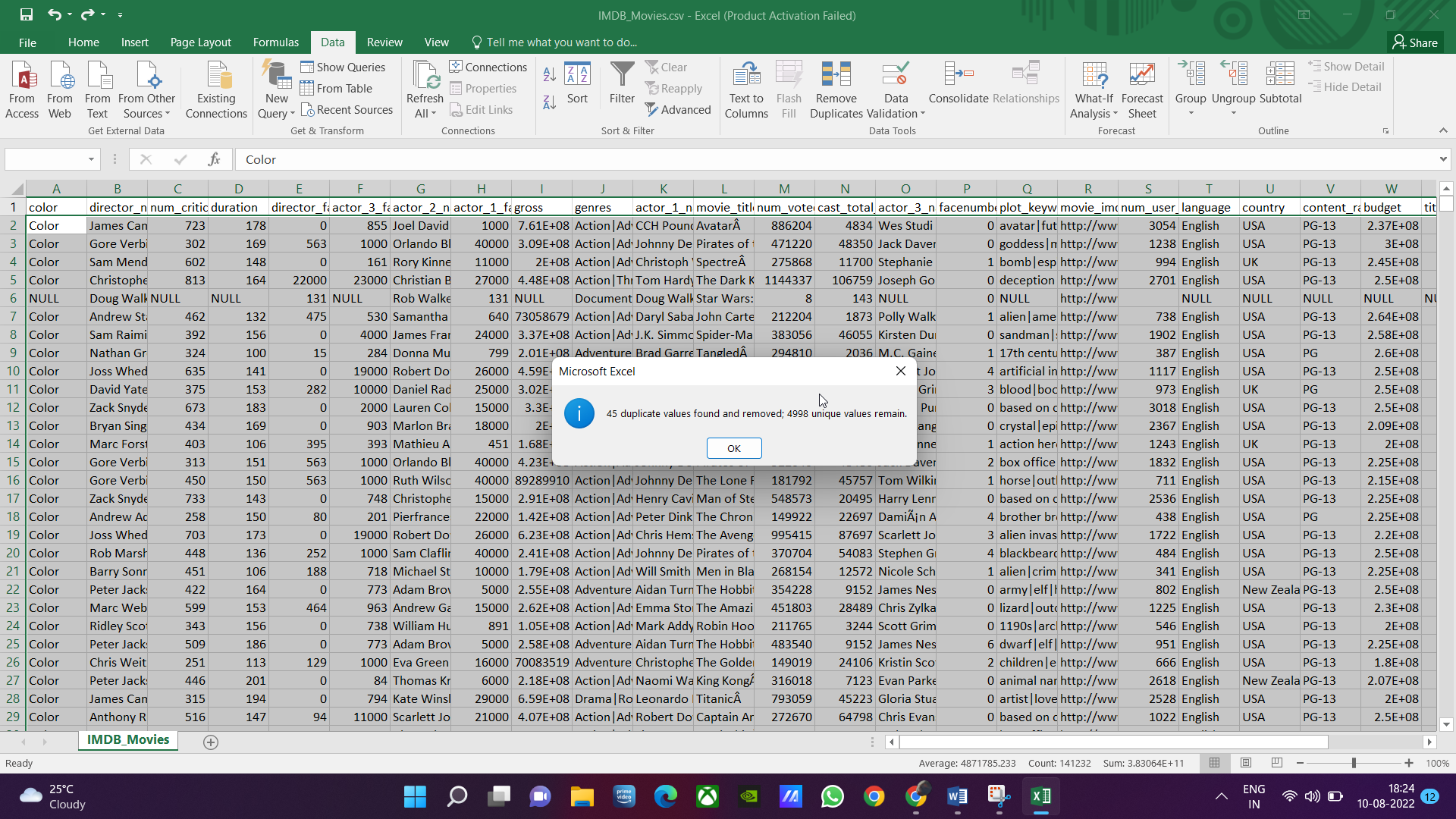
1. Selecting all data
2. go to special
3. click blanks and click ok (all the blank cells in data cell will be selected at the same time)
4. Write null in active cell
5. Hit ctrl+ enter
6. Now all empty cell will be field by null



Removing duplicate rows

1. Select all
2. Go to data
3. Remove duplicates
4. Select my data has headers
5. Click ok

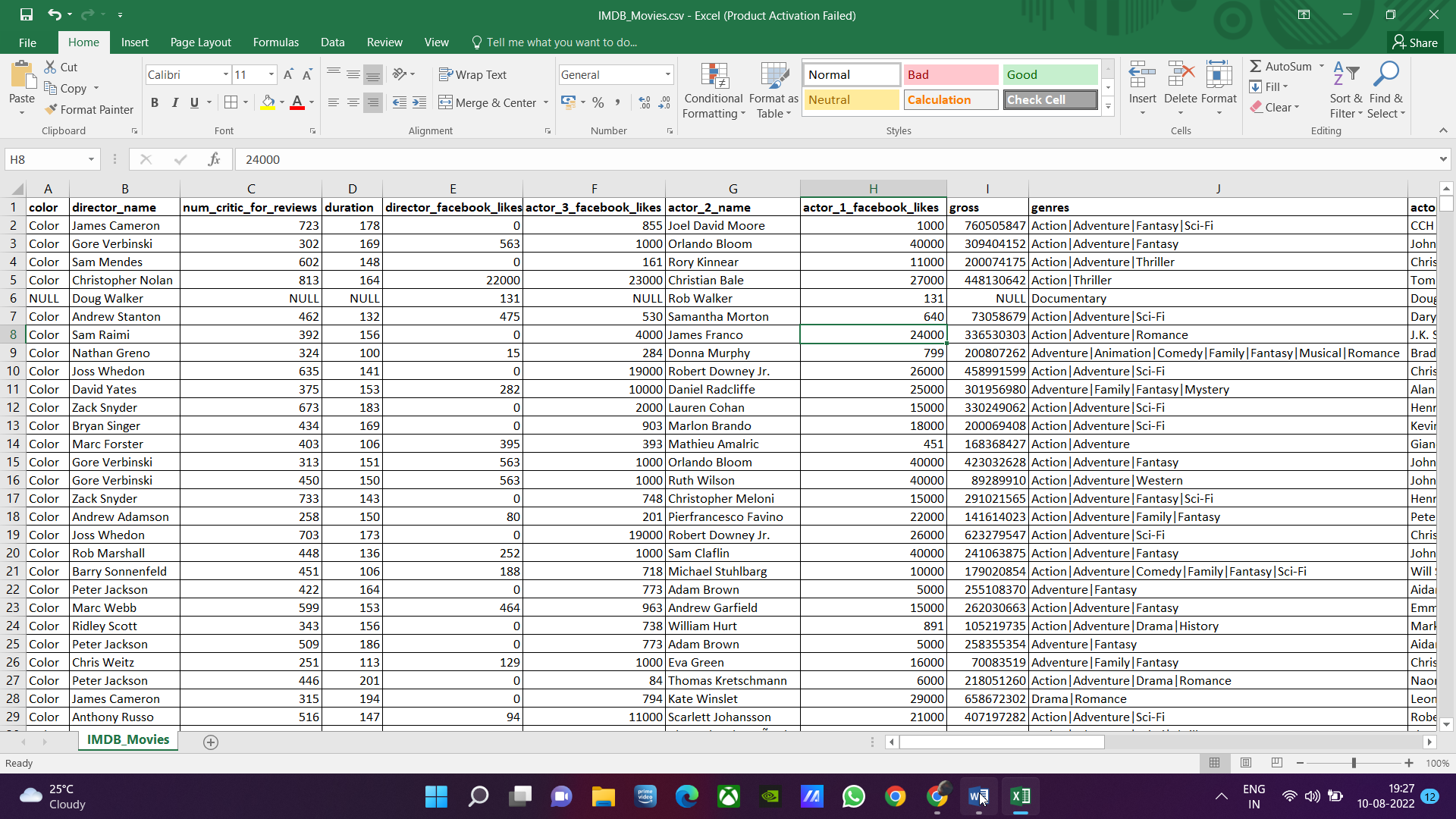


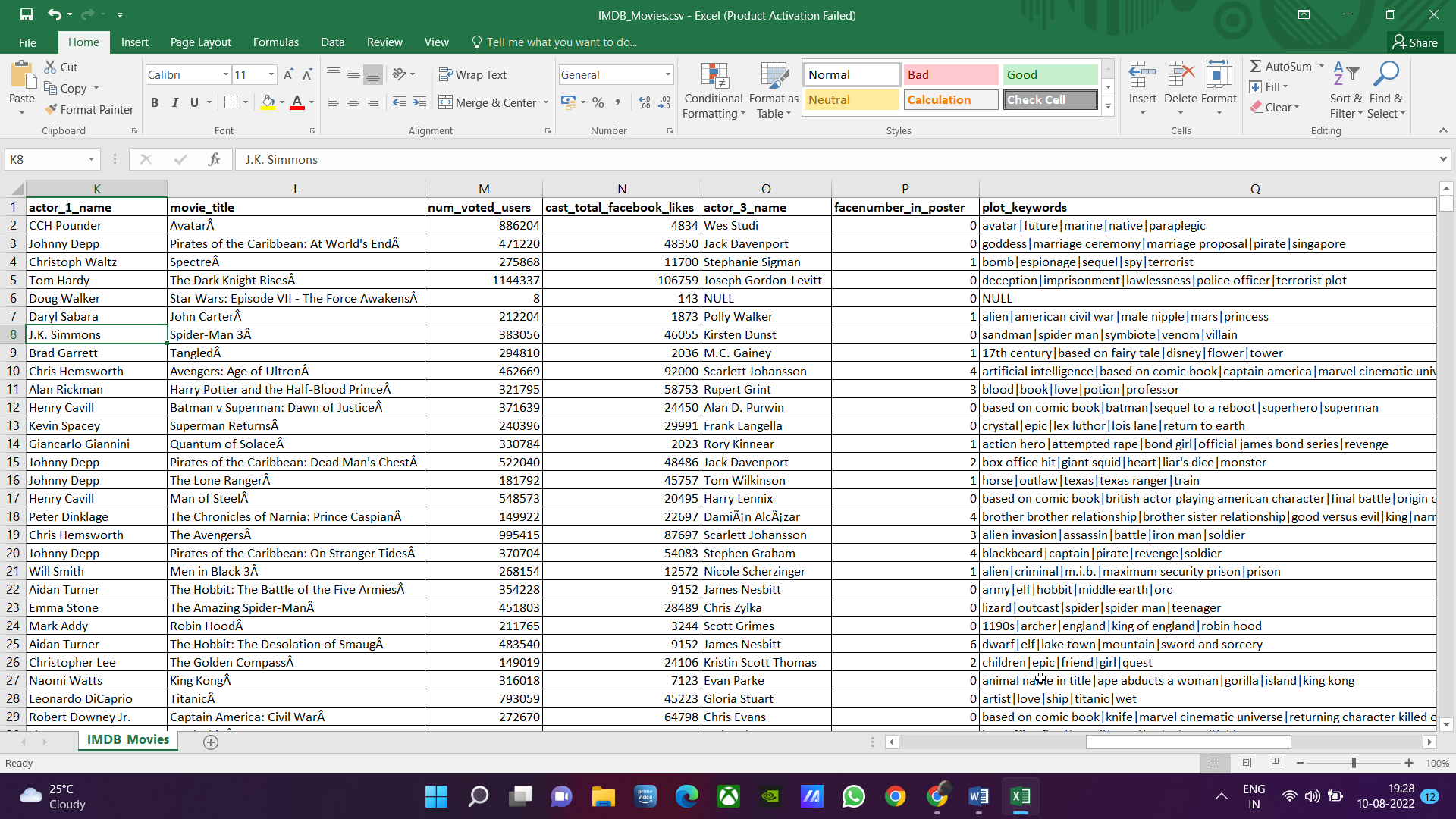
 Highlighting errors:

1. Select all
2. Go to special
3. Select formulas
4. Select only errors from formulas
5. Click ok

Adding borders:

1. Select all
2. Add border
3. Proper spacing to view database properly
4. Making header text bold





1. **Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

**Soln:**

* Movie name: AvatarÂ
* Budget: 237000000
* Gross: 760505847
* Profit: 523505847

1. **Top 250:** Create a new column IMDb\_Top\_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb\_score). Also make sure that for all of these movies, the num\_voted\_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.  
     
   Extract all the movies in the IMDb\_Top\_250 column which are not in the English language and store them in a new column named Top\_Foreign\_Lang\_Film. You can use your own imagination also!

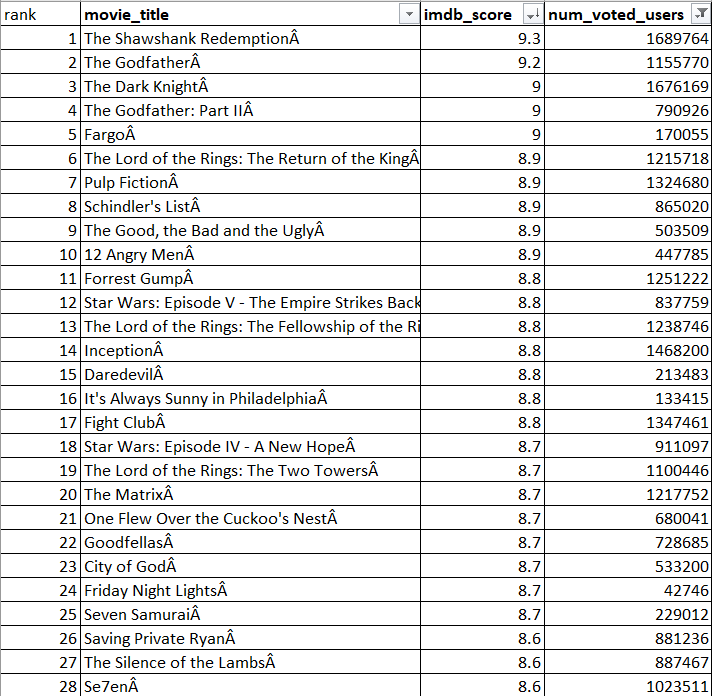
**Soln:**

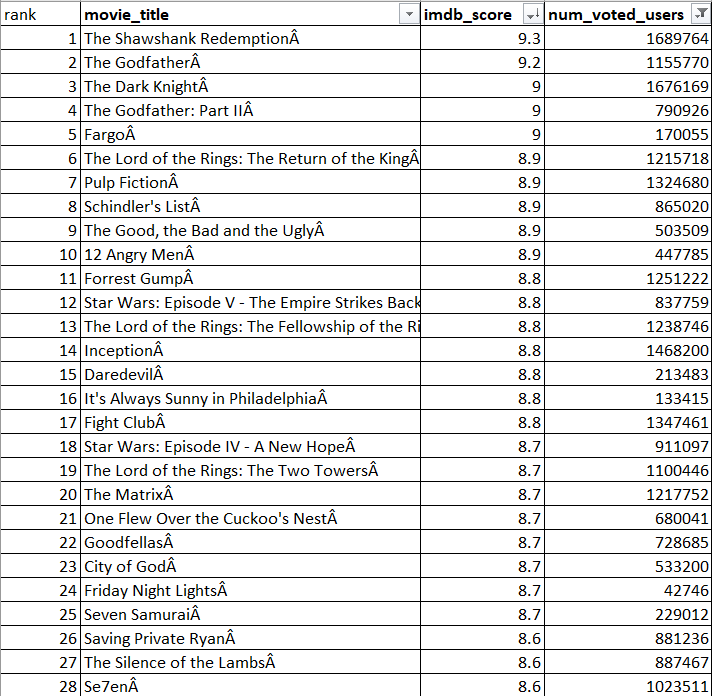
**TOP 250 any language:**

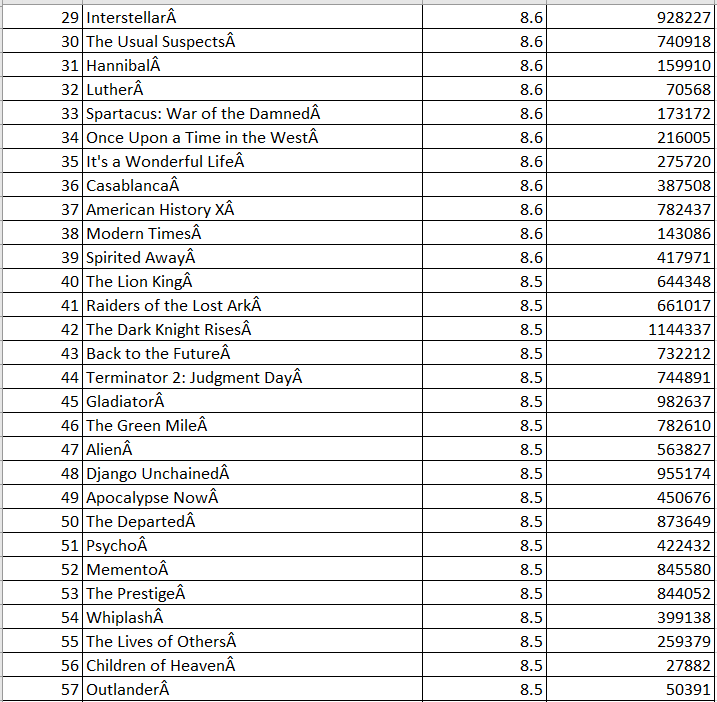
num\_voted\_users: filter >25000

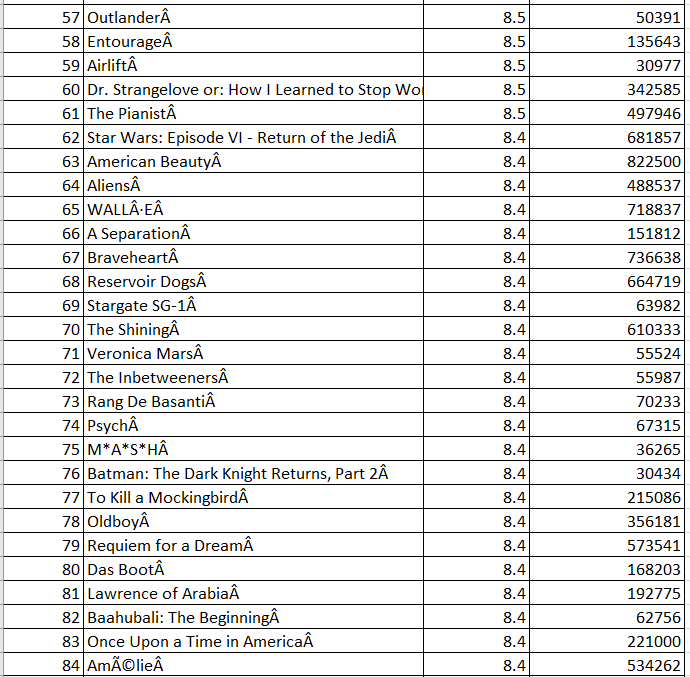
Imbd\_score: sort Largest to smallest, filter top 250

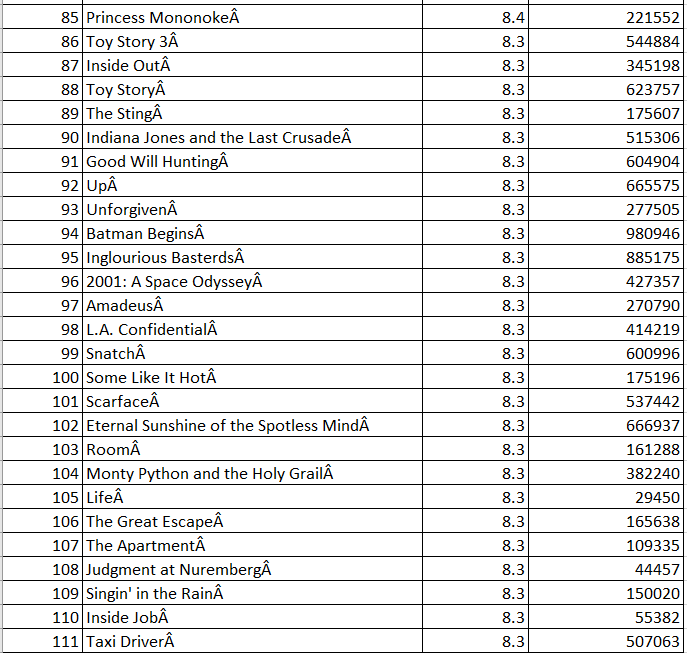
Rank: ROW() – 1

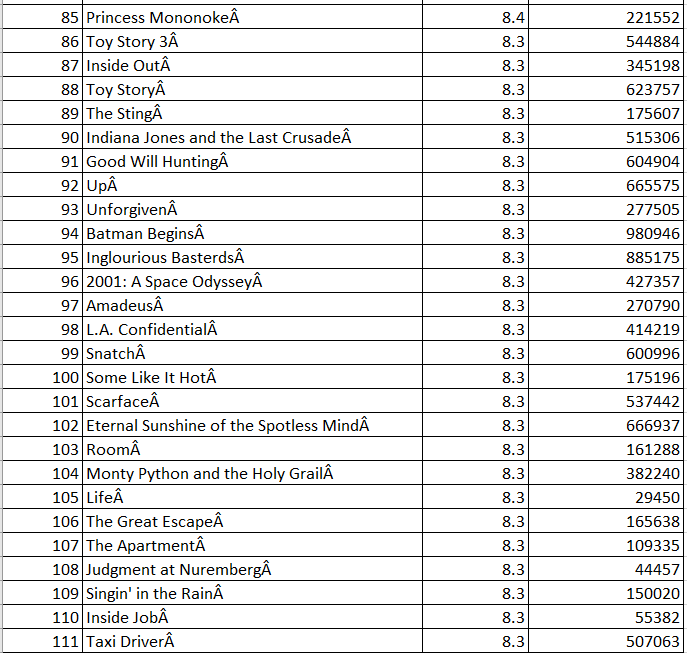


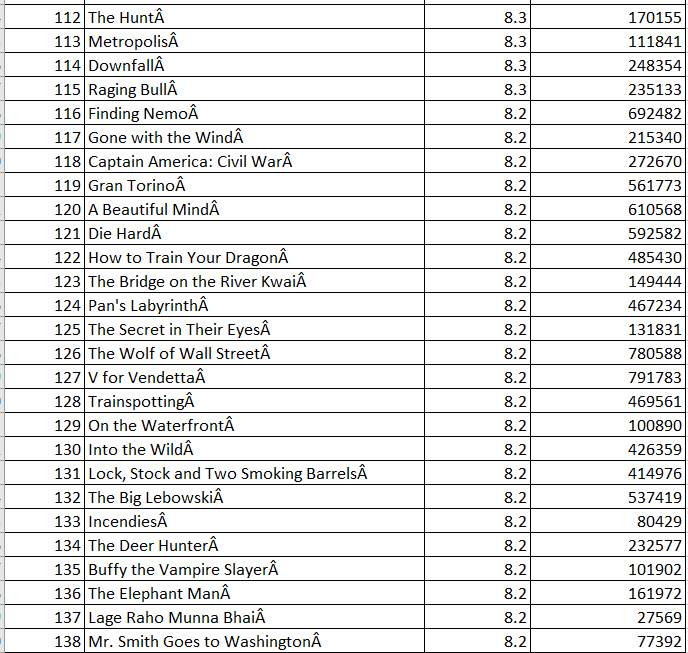


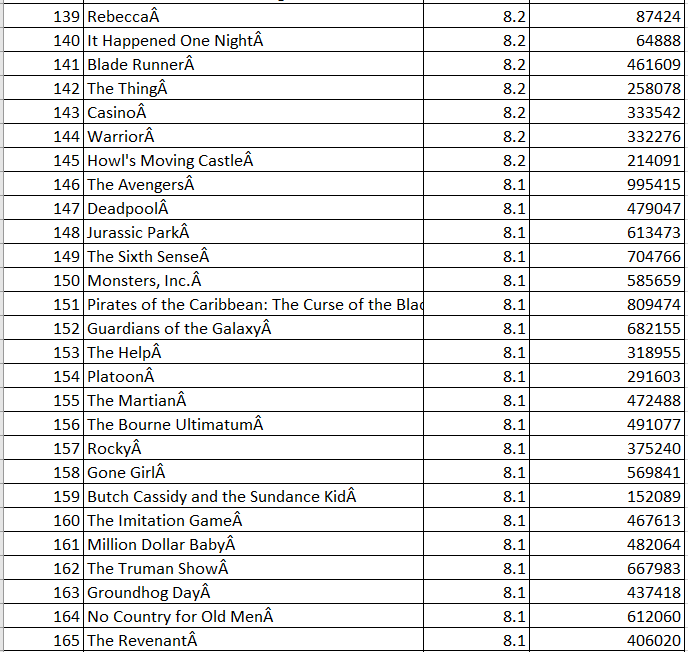


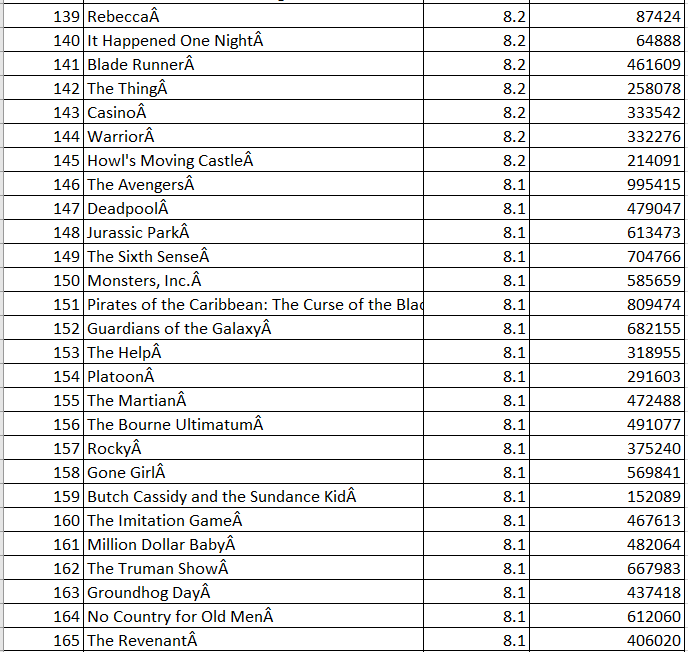


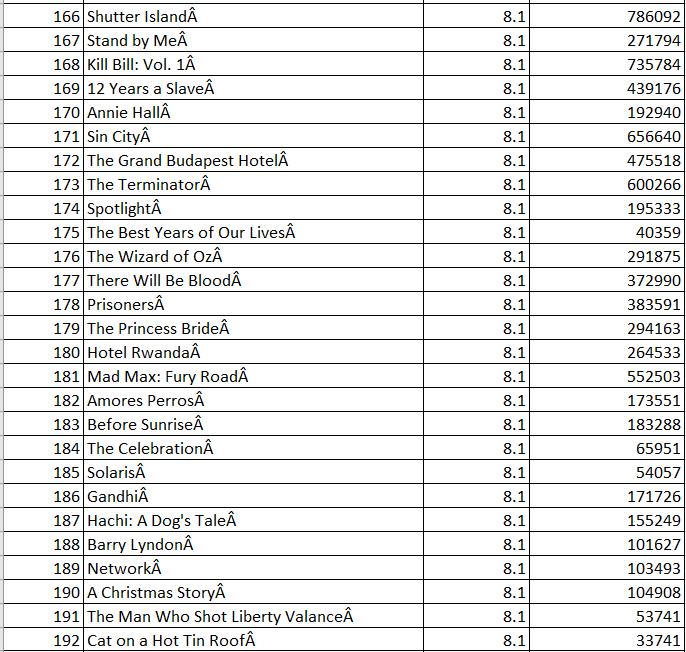


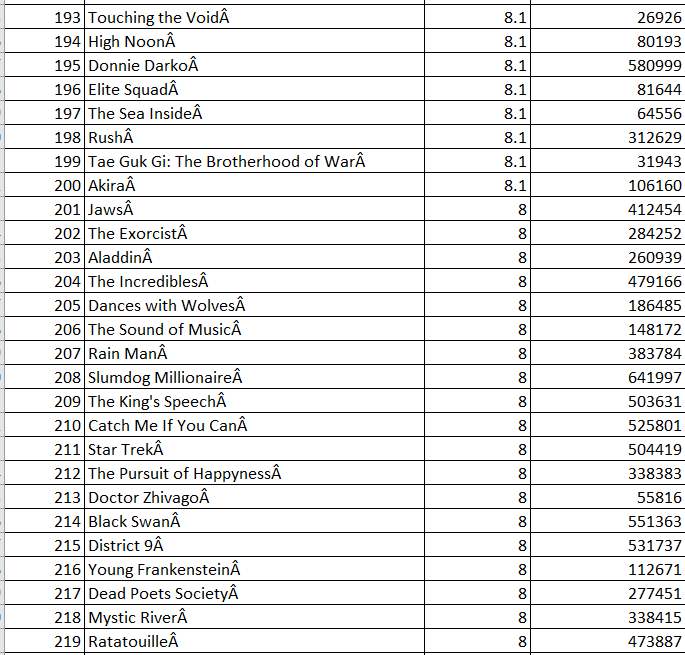


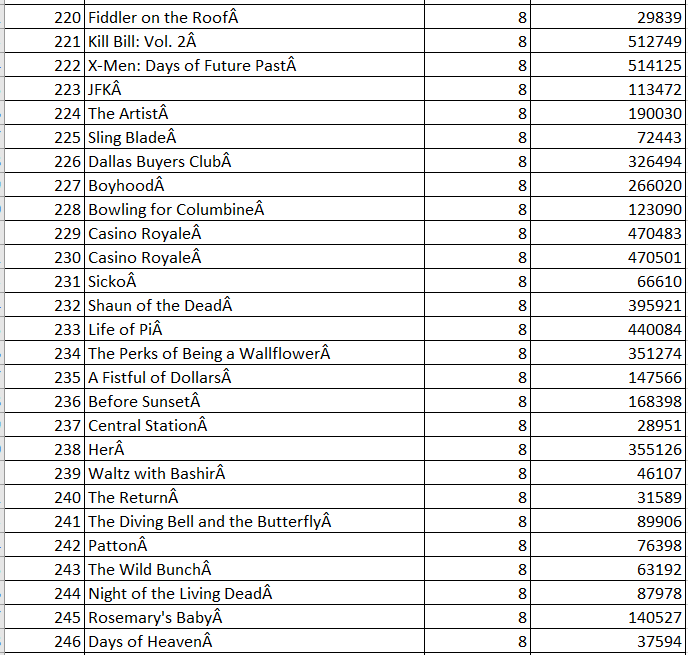


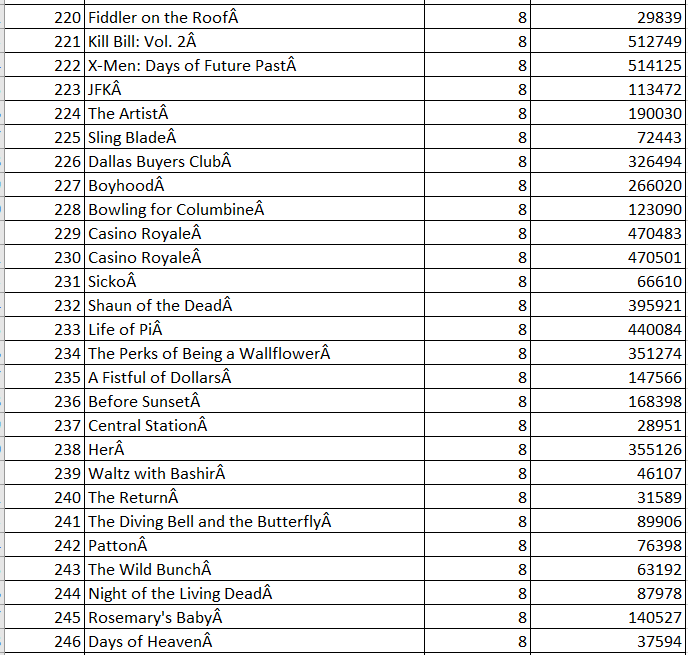






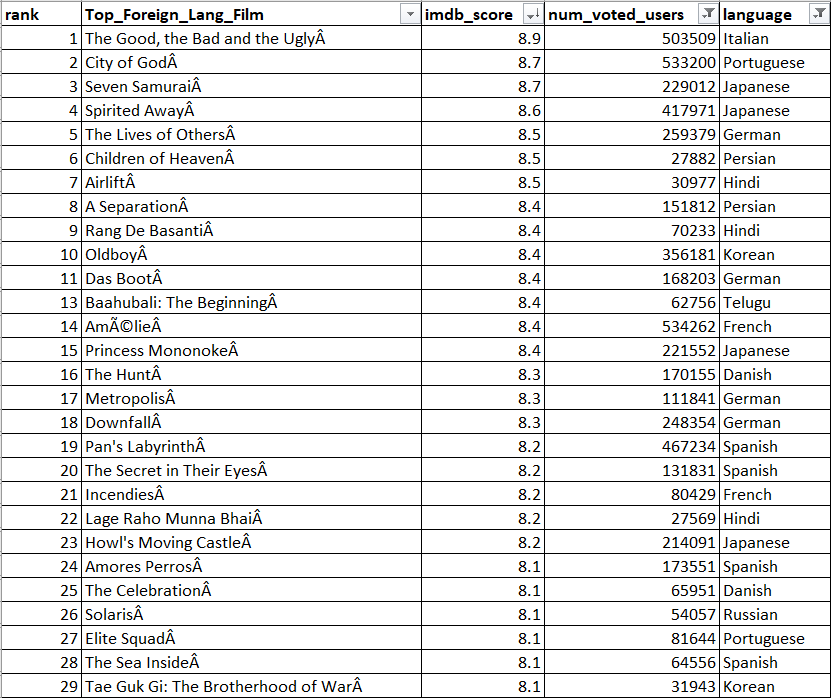


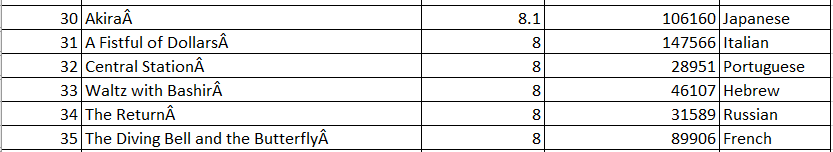






**Top\_Foreign\_Lang\_Film:**

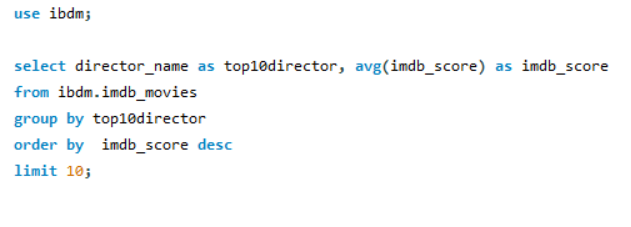




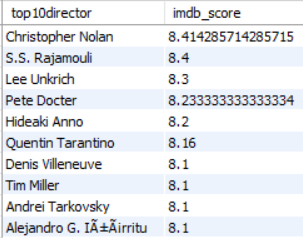
1. **Best Directors:**Group the column using the director\_name column.  
     
   Find out the top 10 directors for whom the mean of imdb\_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.

**Soln:**

**Query:**



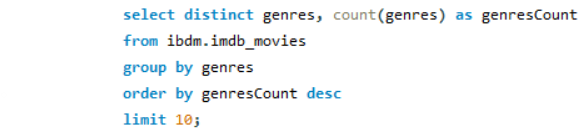
**Result:**



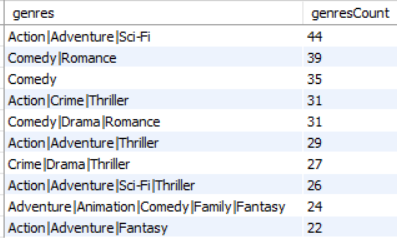
1. **Popular Genres:**Perform this step using the knowledge gained while performing previous steps.

**Soln:**

**Query:**

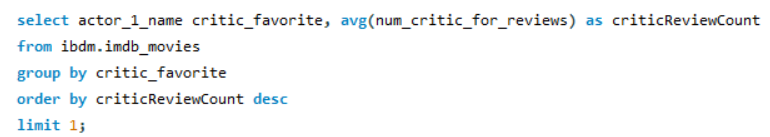


**Result:**

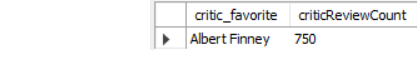


1. **Charts:**Create three new columns namely, Meryl\_Streep, Leo\_Caprio, and Brad\_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor\_1\_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.  
     
   Append the rows of all these columns and store them in a new column named Combined.  
     
   Group the combined column using the actor\_1\_name column.  
     
   Find the mean of the num\_critic\_for\_reviews and num\_users\_for\_review and identify the actors which have the highest mean.  
     
   Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title\_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df\_by\_decade.

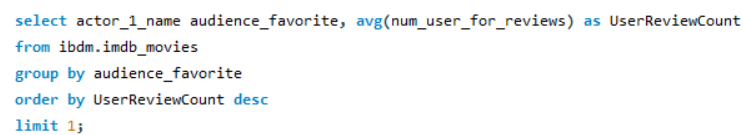
**Query 1: Critic favourite**



**Output:**



**Query 2: Audience Favourite**



**Output:**



**Result:**

* In this project I have gain practical hands on knowledge to analyse numerical data in an Excel spreadsheet.
* Learnt various MS Excel functions and formulas that can be used in many companies in day to day analysis data record in their company