##### PROJECT 5: IMDB Movie Analytics-I

Final Project – I

NAME: K. Yaswanth

GMAIL: [kunayaswanth123@gmail.com](mailto:kunayaswanth123@gmail.com)

To provide the best report for the Internet Movie Database with the help of MS Excel knowledge.

SUBMITTING TO TRAINITY :)

SPECIAL THANKS TO TRAINITY TEAM

**Project Description:**

1. IMDB (Internet Movie Database) is a popular online platform where users can rate and review movies.
2. In this project, we will be analysing a dataset containing various features of movies, such as budget, genre, runtime, and cast, along with their corresponding IMDB ratings.
3. The objective of the project is to explore the IMDb dataset and perform exploratory data analysis

**Approach:**

1. Downloading IMDB\_movies provided. Understanding the data.
2. Clean the data, like removing outliers, removing irrelevant columns, handling missing values, and formatting the data for better analysis.
3. Analyse the dataset using Excel functions and formulas to find meaningful insights such as movies with highest profit, best directors, popular movies etc.

**Tech – Stack Used:**

1. Microsoft Excel
2. Dataset statistics (given)

**Insights:**

Through this project, it became easy to learn the real-life project experiences and learned how to use function in large amount of data and help the imdb company.

The below are questions and answers for this imdb movie analytics are as follows:

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

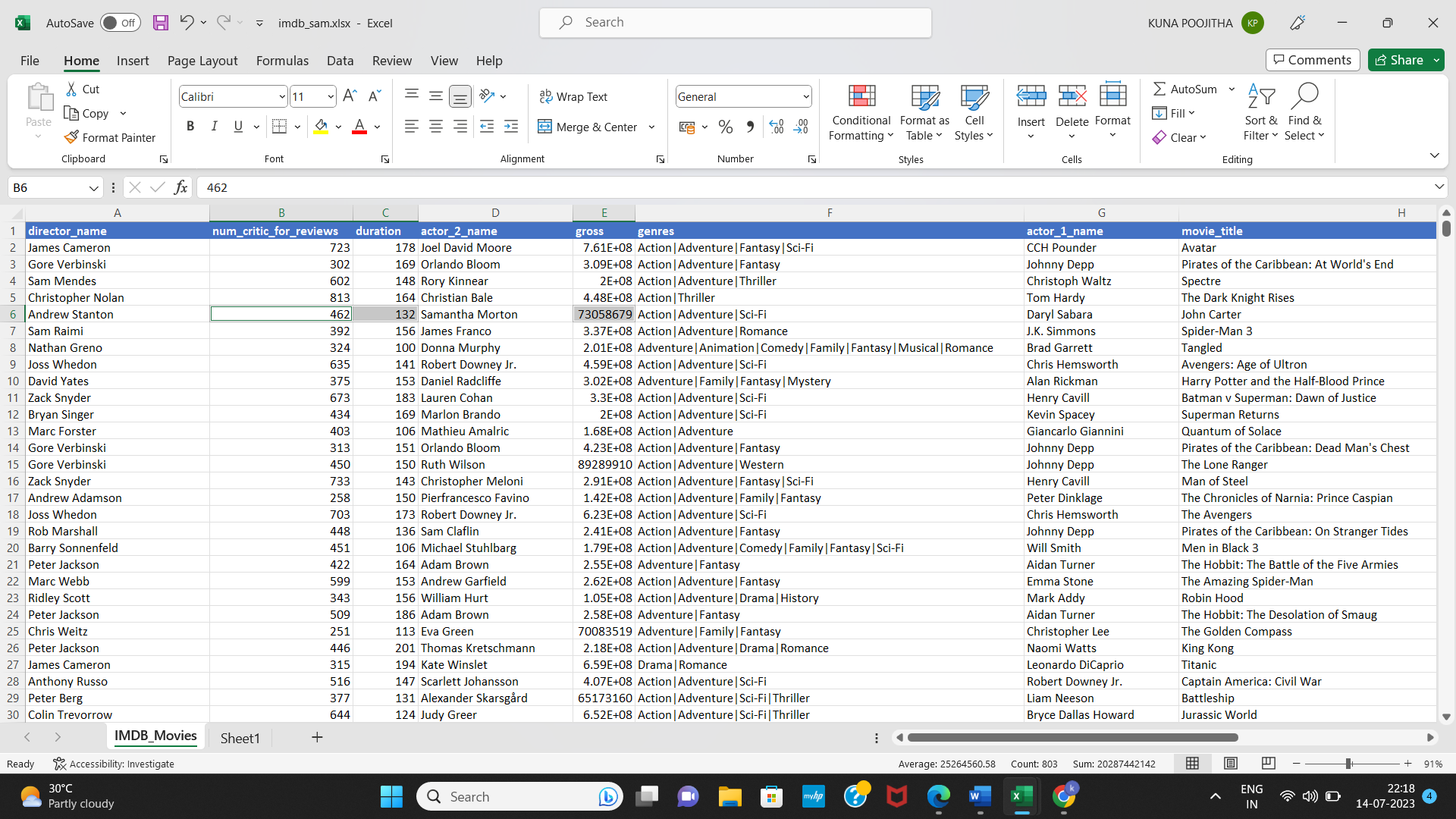
Ans: importing the csv file to MS Excel sheet.

1. Remove the duplicates (By going to data tab and clicking on remove duplicates). All duplicates are removed.
2. Remove unwanted columns
3. Remove the null values which contain the rows.

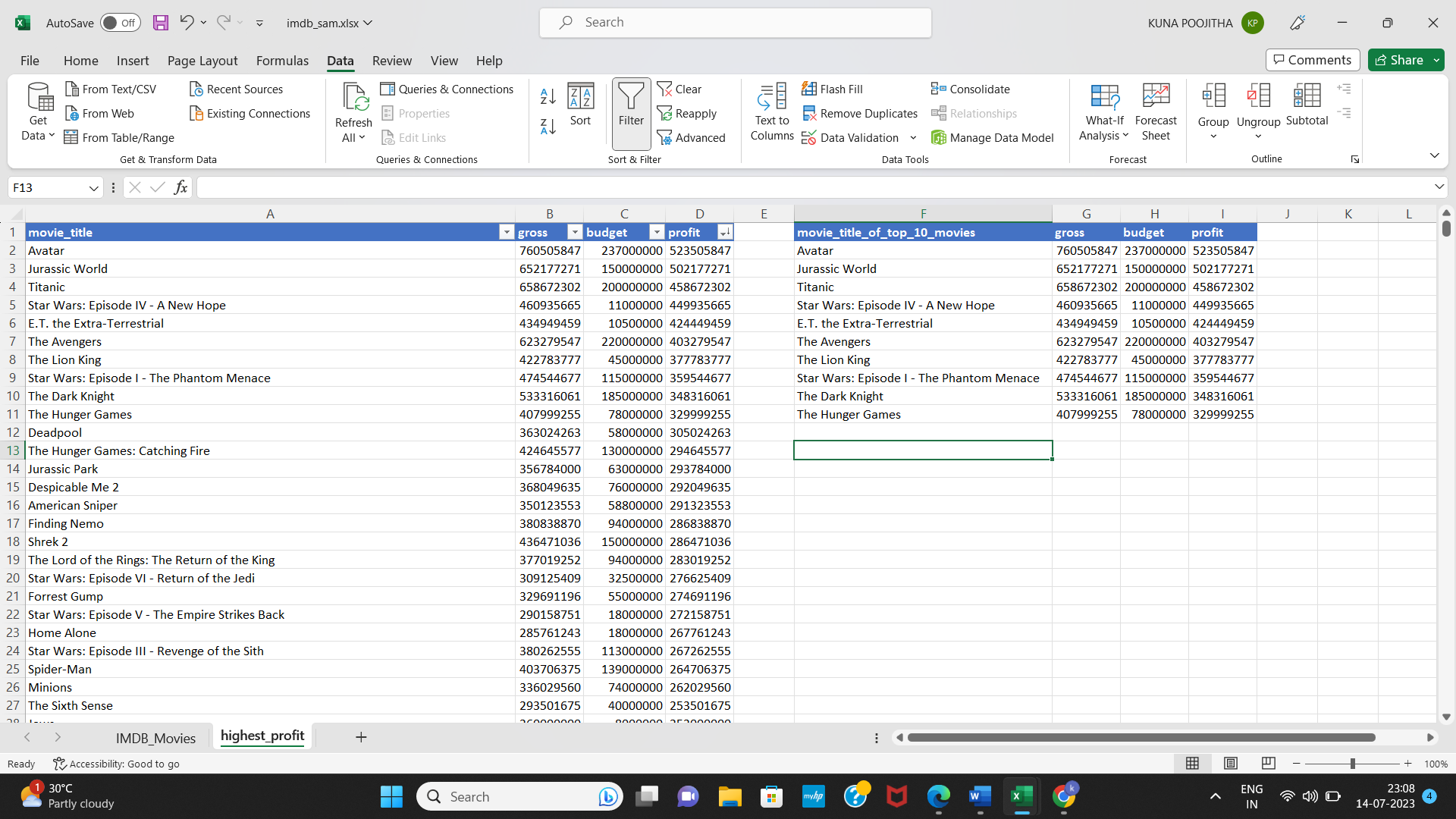
Removing unwanted columns by selecting columns and press ctrl + “- “

To remove null valued row, we go to special and select blank then press ctrl + “- “

In this way we clean the data.



**B. 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.  
**Your task:** Find the movies with the highest profit?

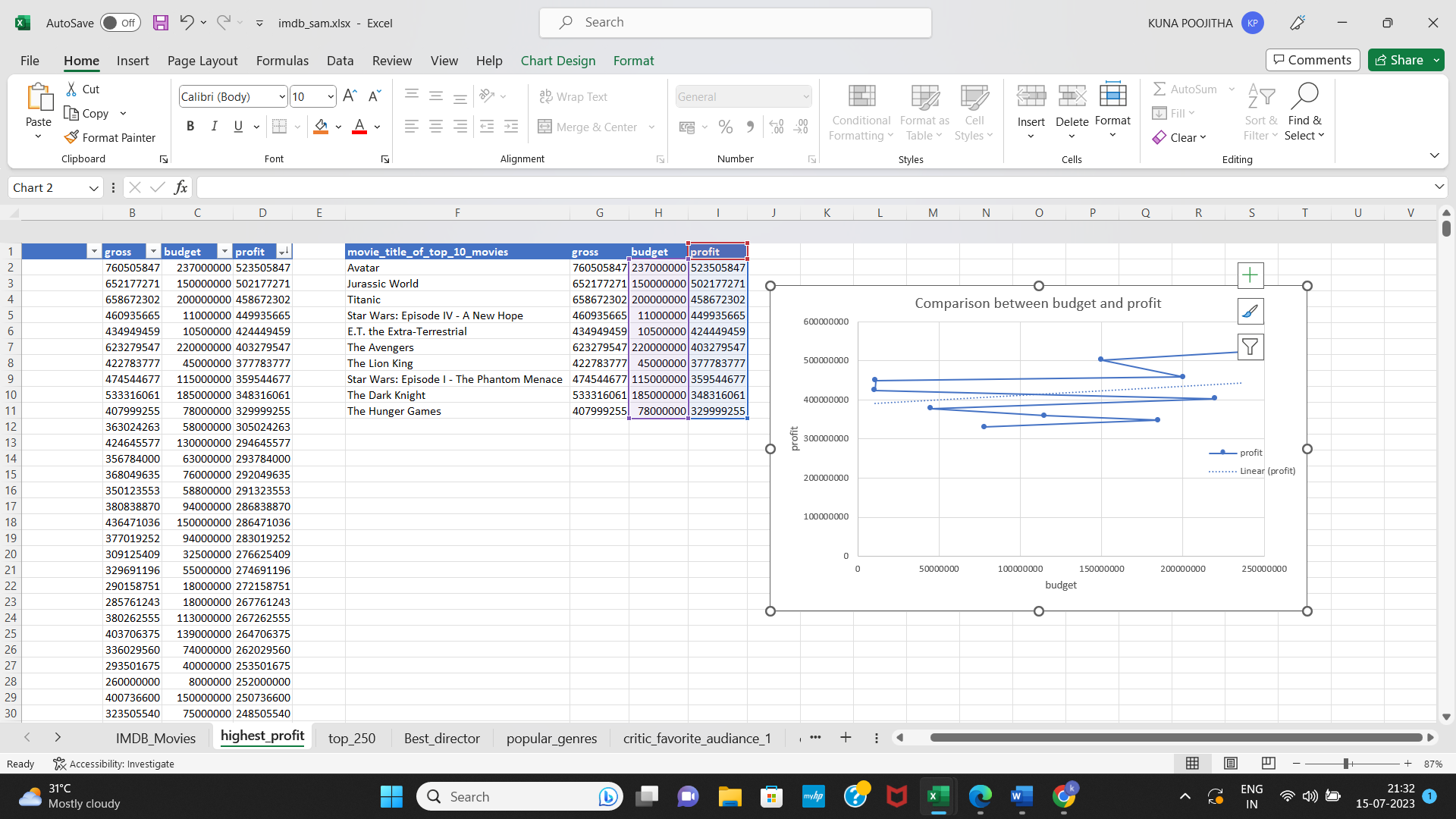


Top 10 movie titles by using filter option in data and we copy the movie\_title column, gross column, budget column and paste it in another worksheet.

Steps:

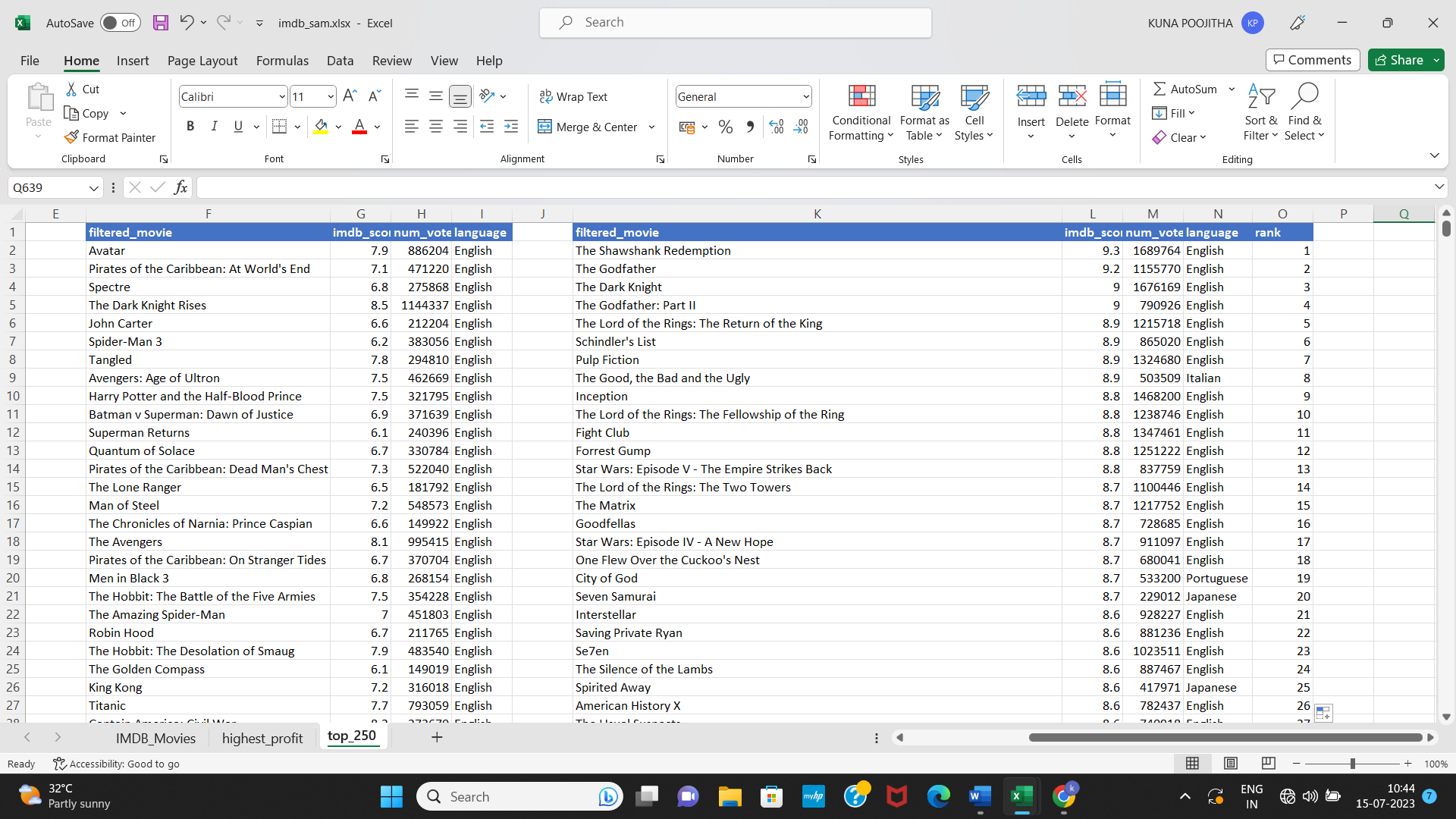
1. Create new column profit which is difference between gross and budget.
2. To find top 10 movies we use resize the table and apply filter operation in data tab in descending order, here we find top 10 columns.

In this way we get the top 10 movies. In descending order.

Chart:

C. **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!  
**Your task:**Find IMDB Top 250

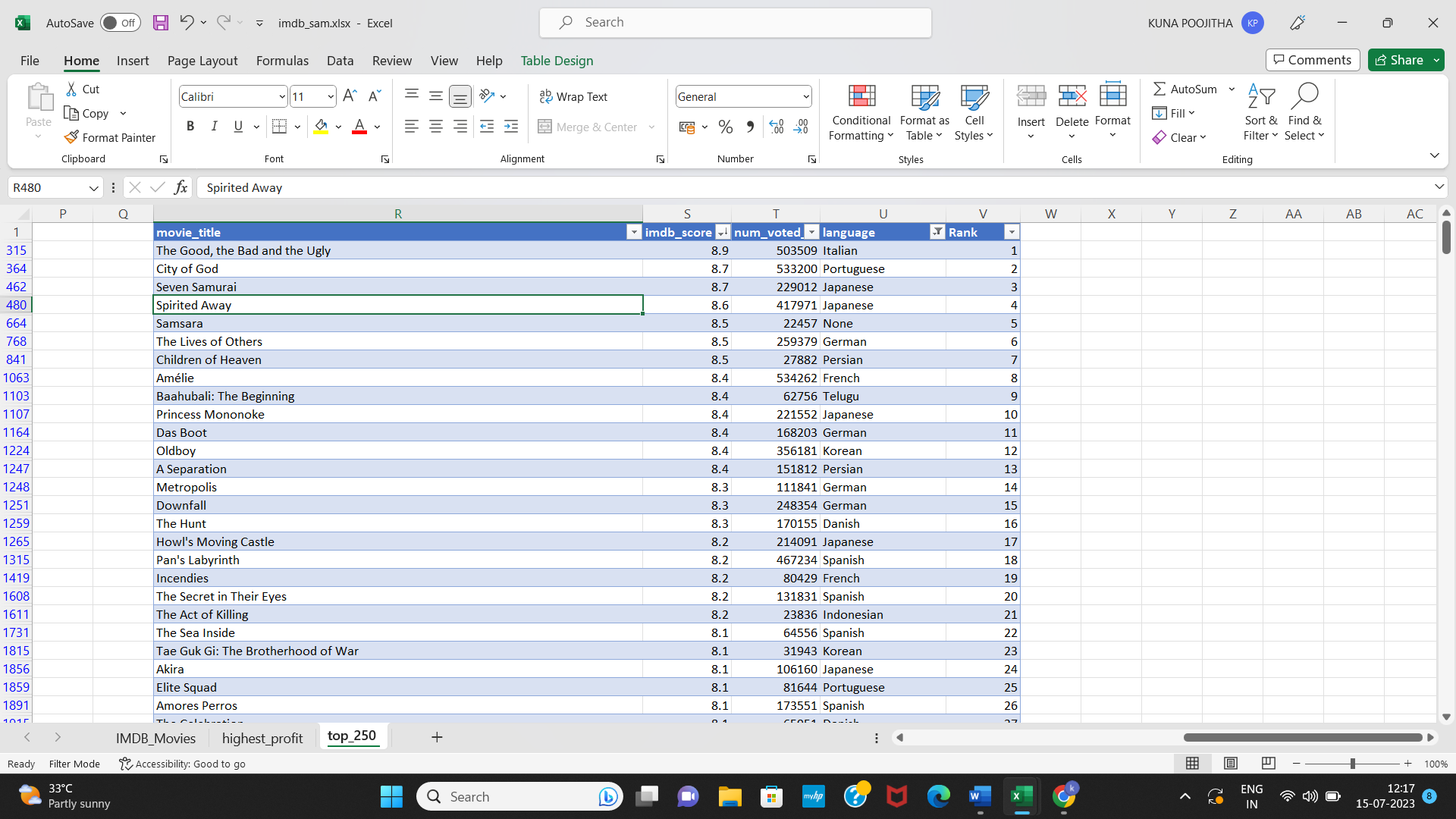
Ans:

Here we filter the num\_voted\_users which is greater than 25,000 using filter function and by using sortby function we use descending order of imdb score.

Steps: -

1. First, filter movies with votes >25000 were found using the FILTER
2. =FILTER(F2:I3856,num\_voted\_users>=25000)
3. To find the top 250 movies
4. =SORT(FILTER(F2:H2633,F2:F2633>=LARGE(F2:F2633,250)),{1,3},{-1,-1})
5. We get top 250 columns

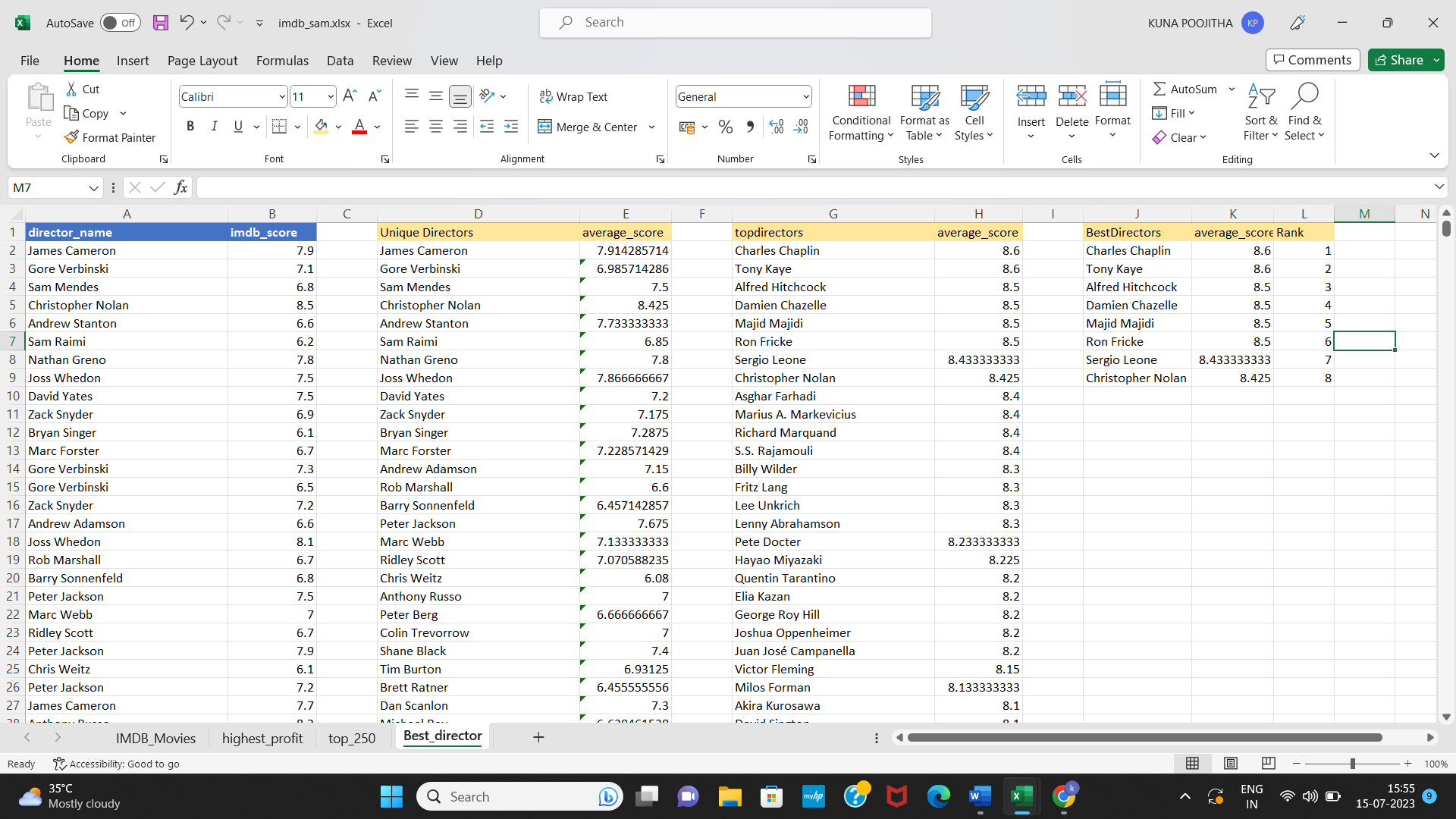
Ans2: To find top 250 in foreign language



Steps:

1. We create the same columns by copying the cells.
2. We create the table, then apply the filter
3. Sort by descending order of imdb\_score
4. Now, we rank the order as 1,2,3…

D. **Best Directors:**To 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.  
**Your task:**Find the best directors

Ans: - 

Steps: -

1. We copy the columns director\_name, imdb\_score from imdb\_movies
2. We need to select all the unique rows. So, that we apply function

=UNIQUE(A2:A3856) to get unique values.

1. We create other column and name as average\_score, then we apply =AVERAGEIF(A2:A3856,D2,B2:B3856) to get all the average values of each director.
2. We sort the directors by applying the function based on requirements is

=SORT(D2:E1752,{2,1},{-1,1})

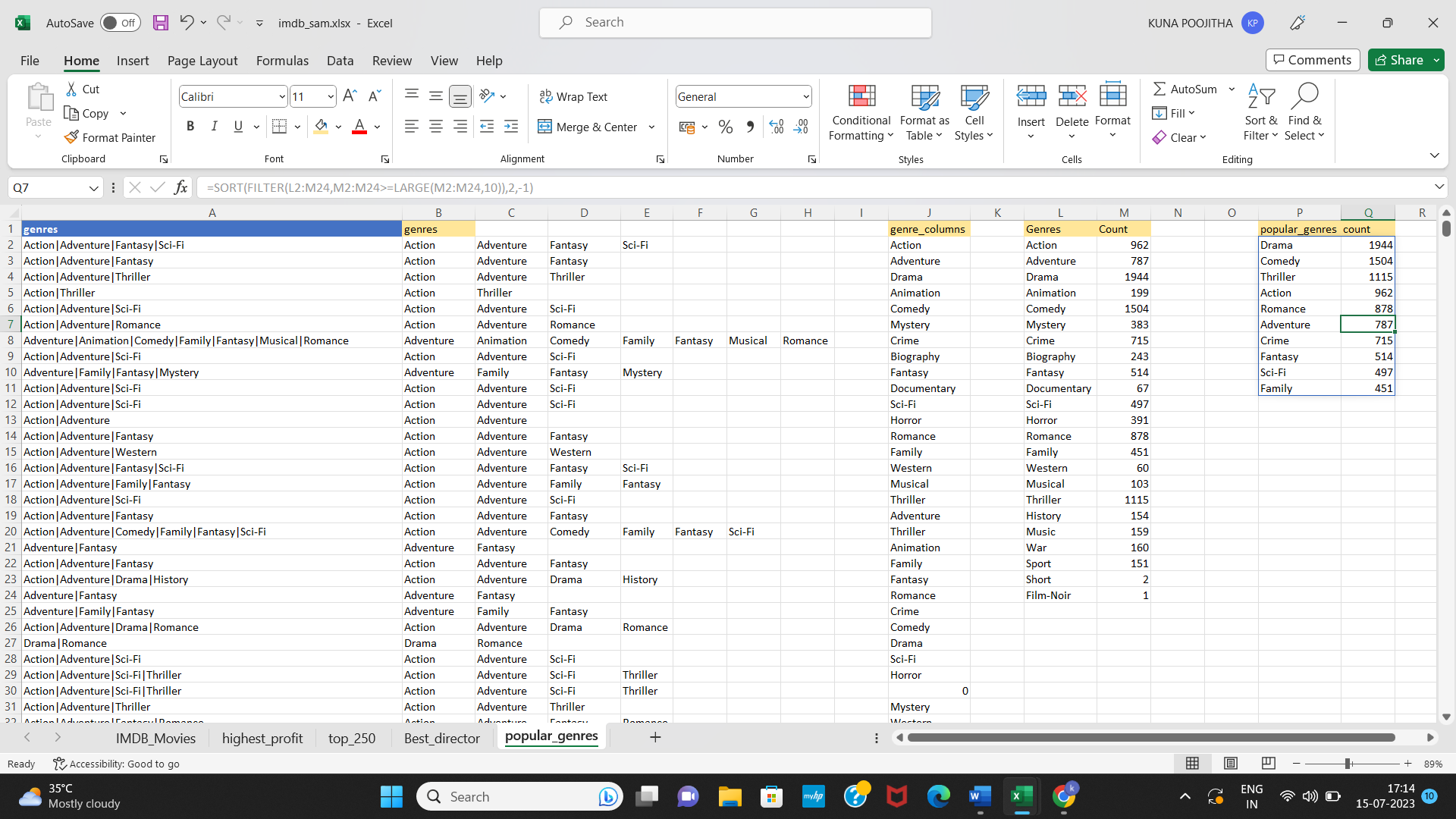
1. We filter the data by applying the filter sort and filter operation for first 10 rows

=SORT(FILTER(D2:E1752,E2:E1752>LARGE(E2:E1752,10)),{2,1},{-1,1})

1. We get desired result.

E. **Popular Genres:**Perform this step using the knowledge gained while performing previous steps.  
**Your task:**Find popular genres

Ans:



Steps: -

1. Copying the genres column in imdb\_movies, then go to data tab and apply text to columns
2. We set all the data to one column
3. Creating other column as genres and apply unique function to get all the rows and exclude 0

=UNIQUE(FILTER(J2:J107,J2:J107<>0))

1. Applying countif function on each entity in genres to get the count

=COUNTIFS($B$2:$H$3856,L2)

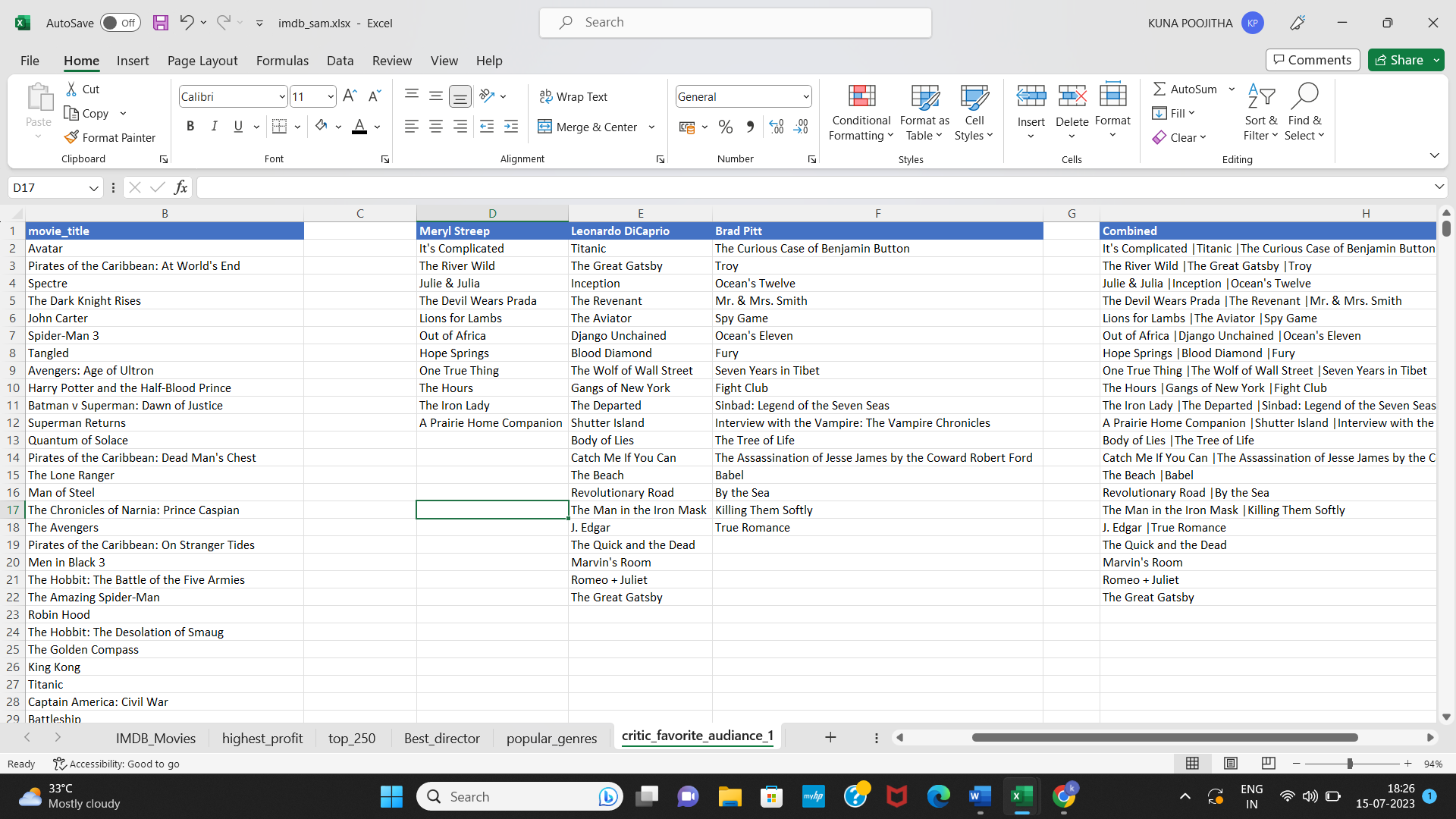
Here $ is fixed.

1. To, find top 10 rows then, by applying sort

=SORT(FILTER(L2:M24,M2:M24>=LARGE(M2:M24,10)),2,-1)

1. We, get desired result.

F. **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.  
  
**Your task:**Find the critic-favorite and audience-favorite actors

Ans 1: -

Steps: -

1. Copy movie\_title, actor\_1\_name from imdb\_movies
2. To filter movies of Meryl Streep,Leonardo DiCaprioand BradPitt

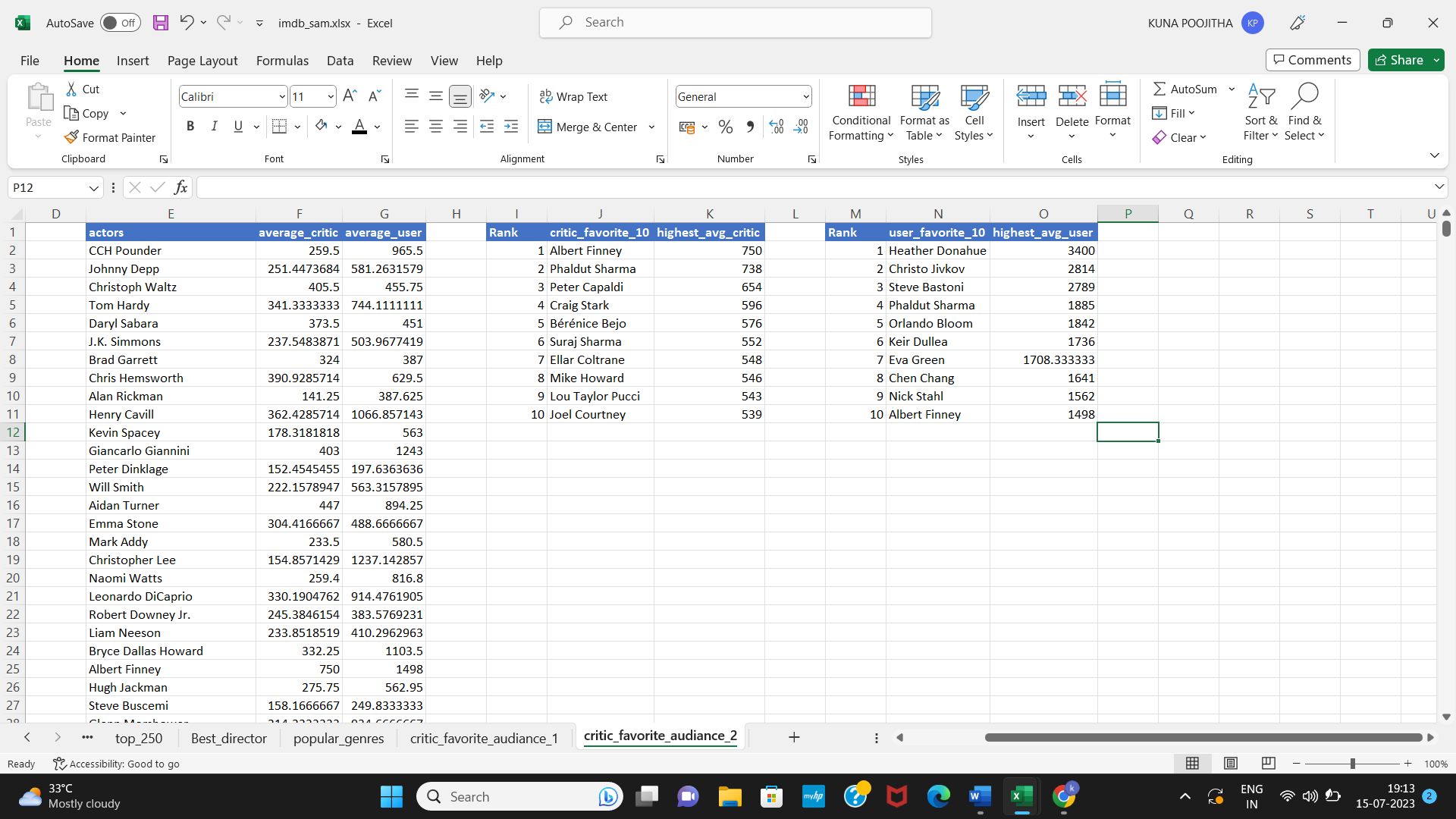
Meryl Streep =FILTER(B2:B3856,A2:A3856=D1)

Leonardo Dicaprioand =FILTER(B2:B3856,A2:A3856=E1)

Brad Pitt =FILTER(B2:B3856,A2:A3856=F1)

1. We get all the movies of the actors which they acted.
2. We join all the movies by using text join using delimiter “|”

=TEXTJOIN("|",TRUE,D2,E2,F2)

Ans\_2:

Steps: -

1. We copy the rows of num\_critic\_for\_reviews , num\_user\_for\_reviews, actor\_1\_name.
2. Filter unique actors using =UNIQUE(C2:C3856)
3. Then we find the average critcs and average users, of each actors.

=AVERAGEIF($C$2:$C$3856,E2,$A$2:$A$3856)

=AVERAGEIF($C$2:$C$3856,E2,$B$2:$B$3856)

1. From this created columns, we create critic\_favourite\_10 and highest\_avg\_critic using

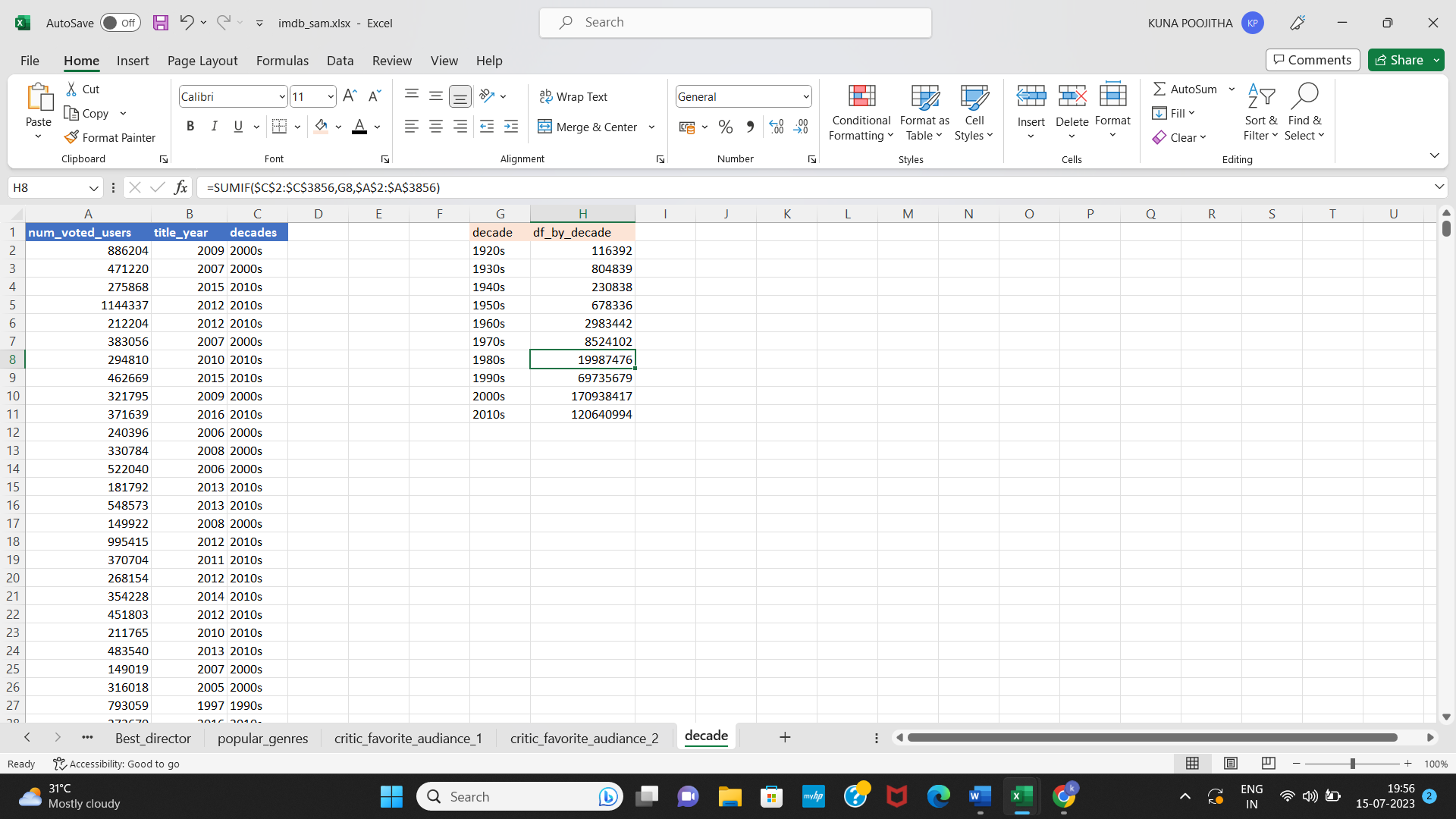
Sort method, getting top 10 critic actors and their average

=SORT(FILTER(E2:F1509,F2:F1509>=LARGE(F2:F1509,10)),2,-1)

1. Similarly, we apply filter operation finding best actor accord to user using

=FILTER(SORT(FILTER(E2:G1509,G2:G1509>=LARGE(G2:G1509,10)),3,-1),{1,0,1})

Filters the top 10 rows according to user, sorting in descending order.

Ans\_3:

To find in a decade total number of voted users.

Steps:

1. Copy the columns num\_voted\_users, title\_year
2. Create decades column and apply operation =LEFT(B2,3)&”0s” to get decades
3. Now, Create decade column and df\_by\_decade
4. In decade we apply =SORT(UNIQUE(C2:C3856),1,1) to get sorted and unique value
5. In df\_by\_decade column we apply =SUMIF($C$2:$C$3856,G2,$A$2:$A$3856)

To get sum of all the num\_voted\_users in each decade.

In this way, we get the critic\_favourite and audience\_favourite actors and corresponding results.

**Excel Drive Link: -** [**https://docs.google.com/spreadsheets/d/1TS1fGRx0MdZlIWz34VAHE5uo3Eafqja6/edit?usp=sharing&ouid=103431433403818797511&rtpof=true&sd=true**](https://docs.google.com/spreadsheets/d/1TS1fGRx0MdZlIWz34VAHE5uo3Eafqja6/edit?usp=sharing&ouid=103431433403818797511&rtpof=true&sd=true)

**Results:** With the help of the project learned how to use functions in formula bar and learned how to use the functions in appropriate manner. Here, we find the best answers for all the questions will definitely benefit the company.

Thank you trainity 😊