**IMDB Movie Analysis**

**Final Project-1**

**Description:**

For the Final Project - 1, we have a dataset with various columns of different IMDB Movies. We are required to Frame the problem. For this task, we will need to define a problem we want to shed some light on.

**Project Approach Used:**

This project is quite challenging and different from the type of project I have worked on previously provided by the team. I am very happy to work on this project and finish it by bringing out insights that will be useful for the company to make better decisions.

**Tech Stack Used:**

In this project, I used

1. **Python,**
2. **Google Collab and**
3. **MS Excel**

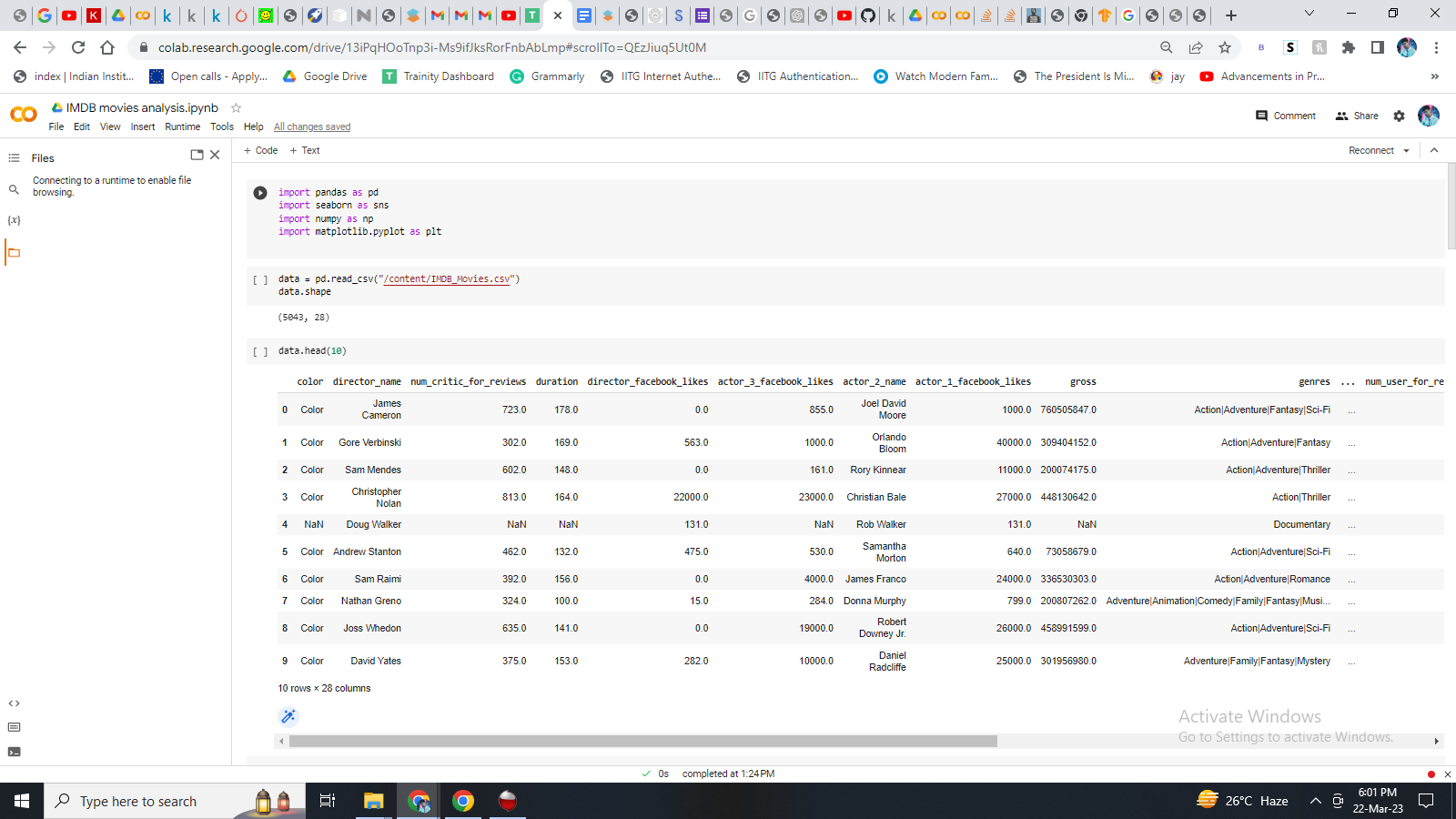
To solve the given problems.

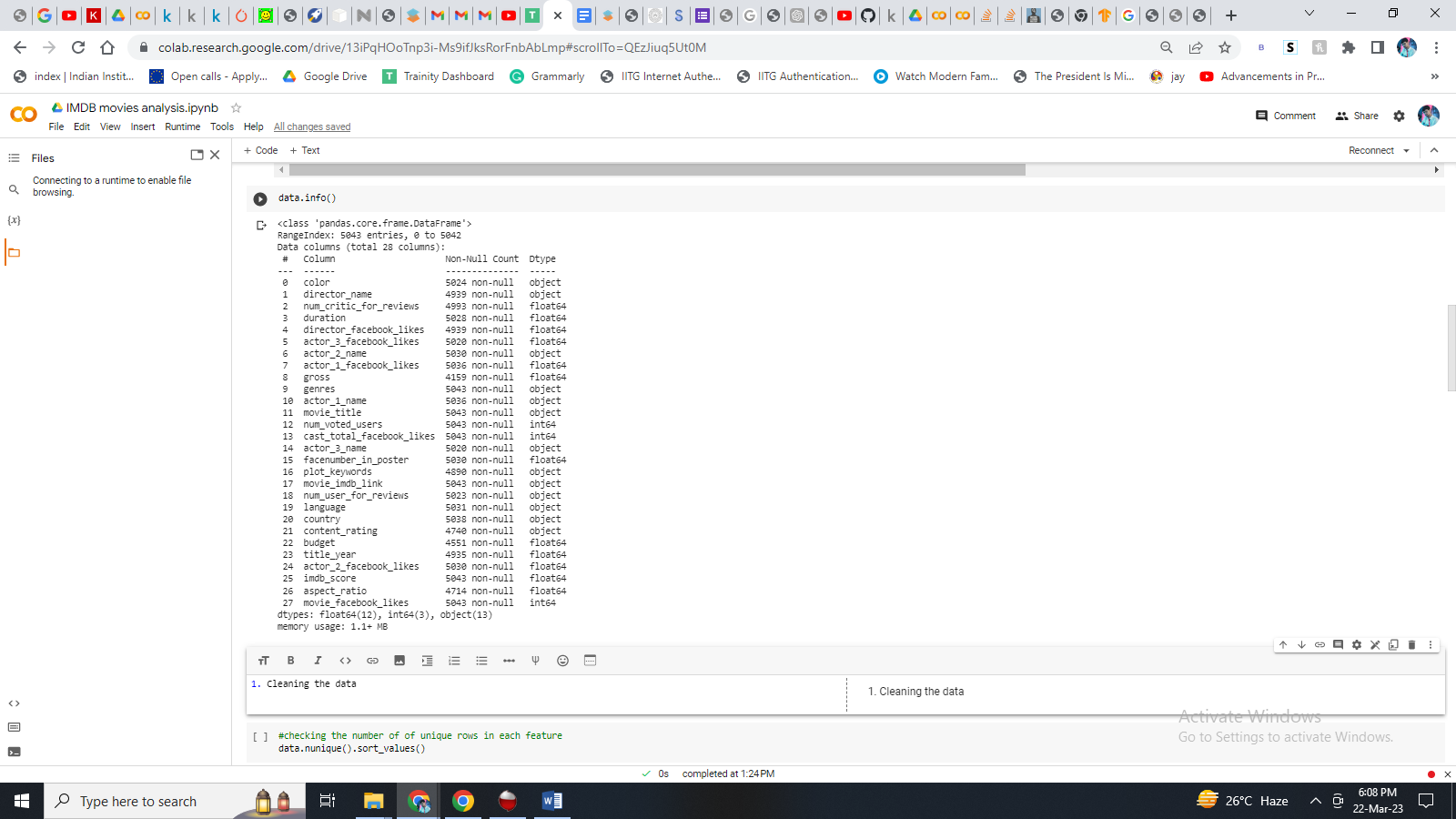
In this project, I achieved some new things like how to get results from huge amounts of data.

The dataset provided by the team has various columns of different IMDB movies. First I started working on the 5 WHY aspect of the dataset.

1. **Cleaning the Data**

**First, we start by exploring the dataset.**



Then we tried to find out some information regarding the dataset.

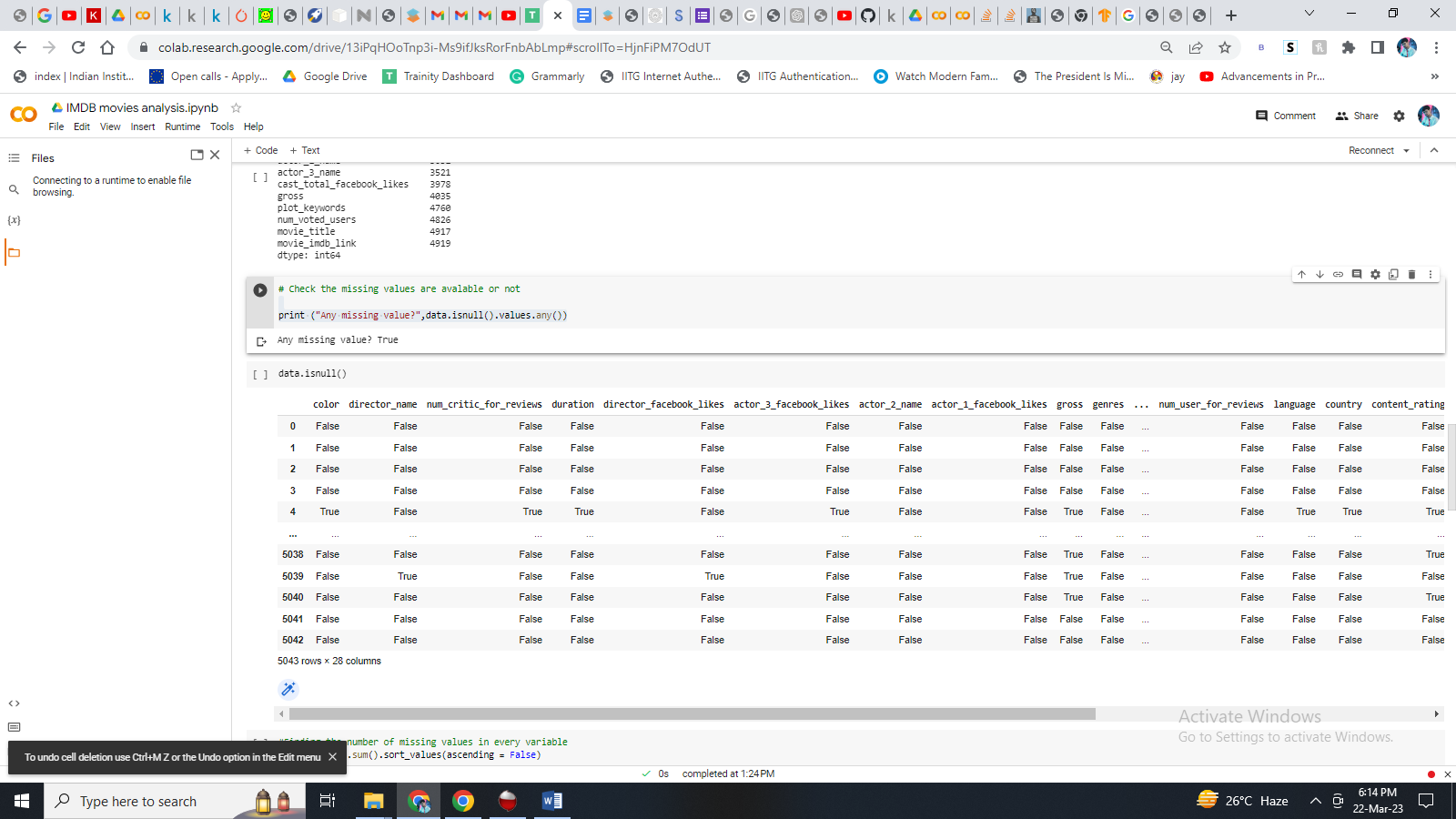
Using – data.info() command.

Then we tried to find out the number of unique rows in each feature.

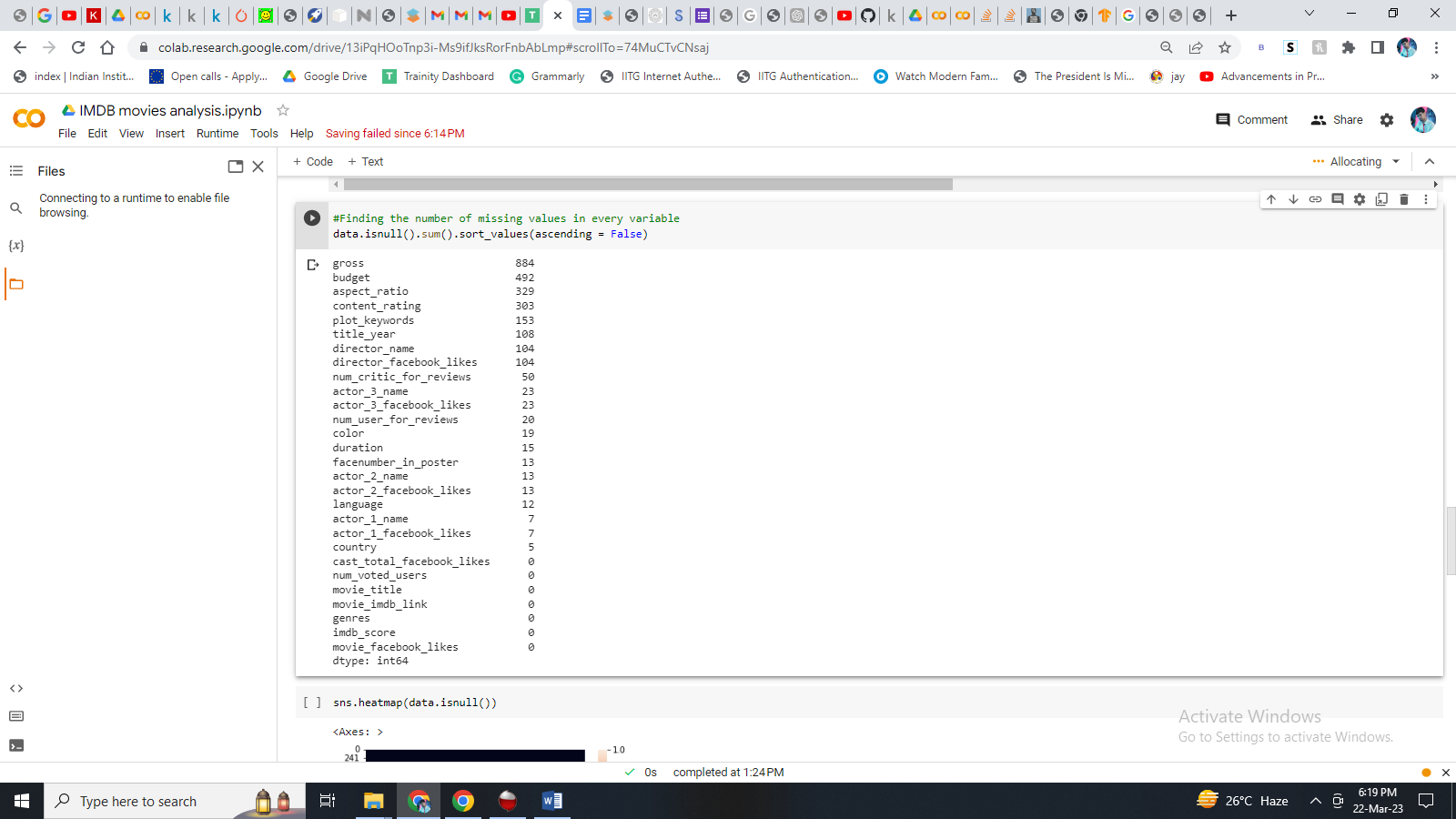
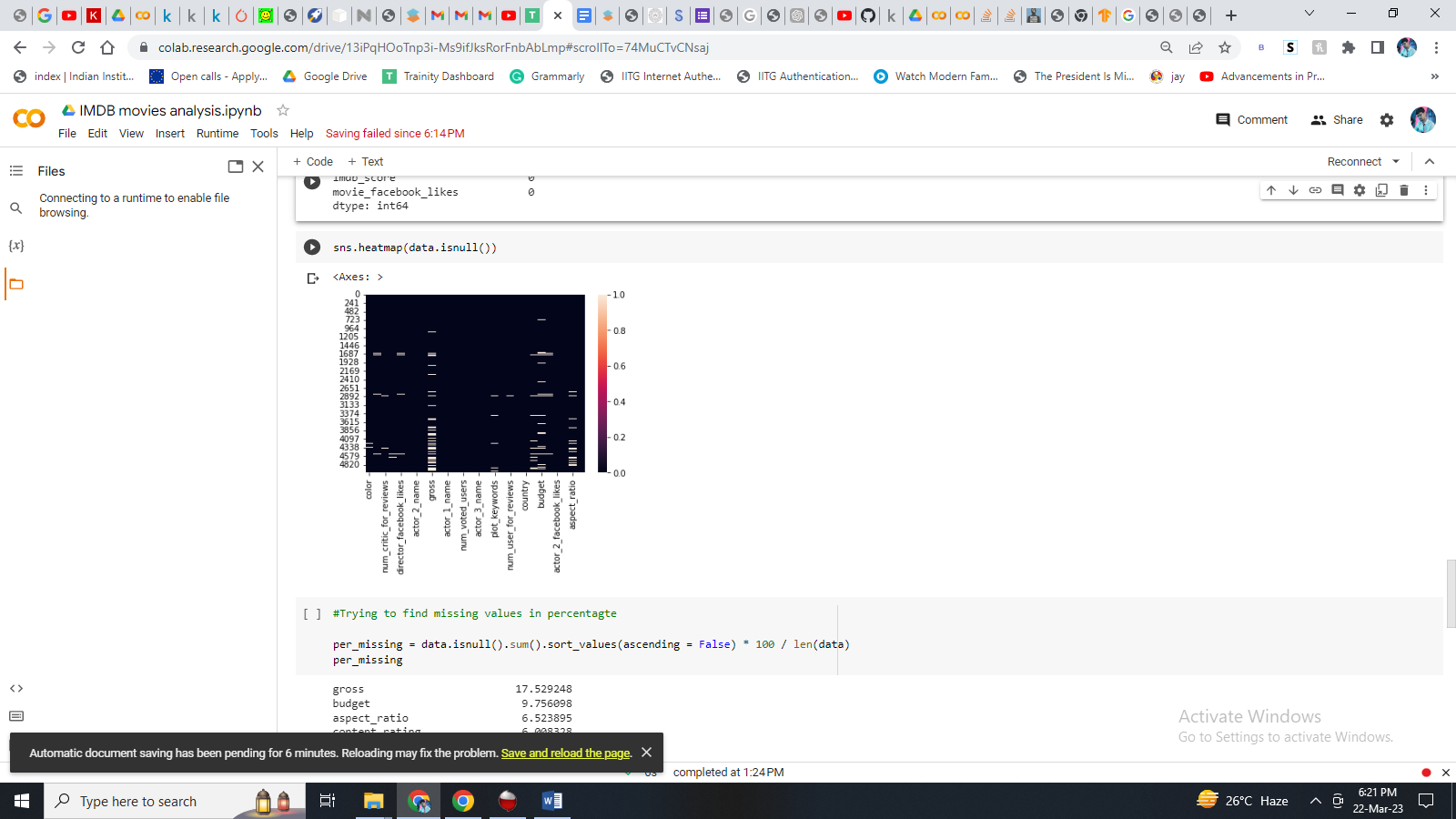
Using – data.nunique().sort\_values() command

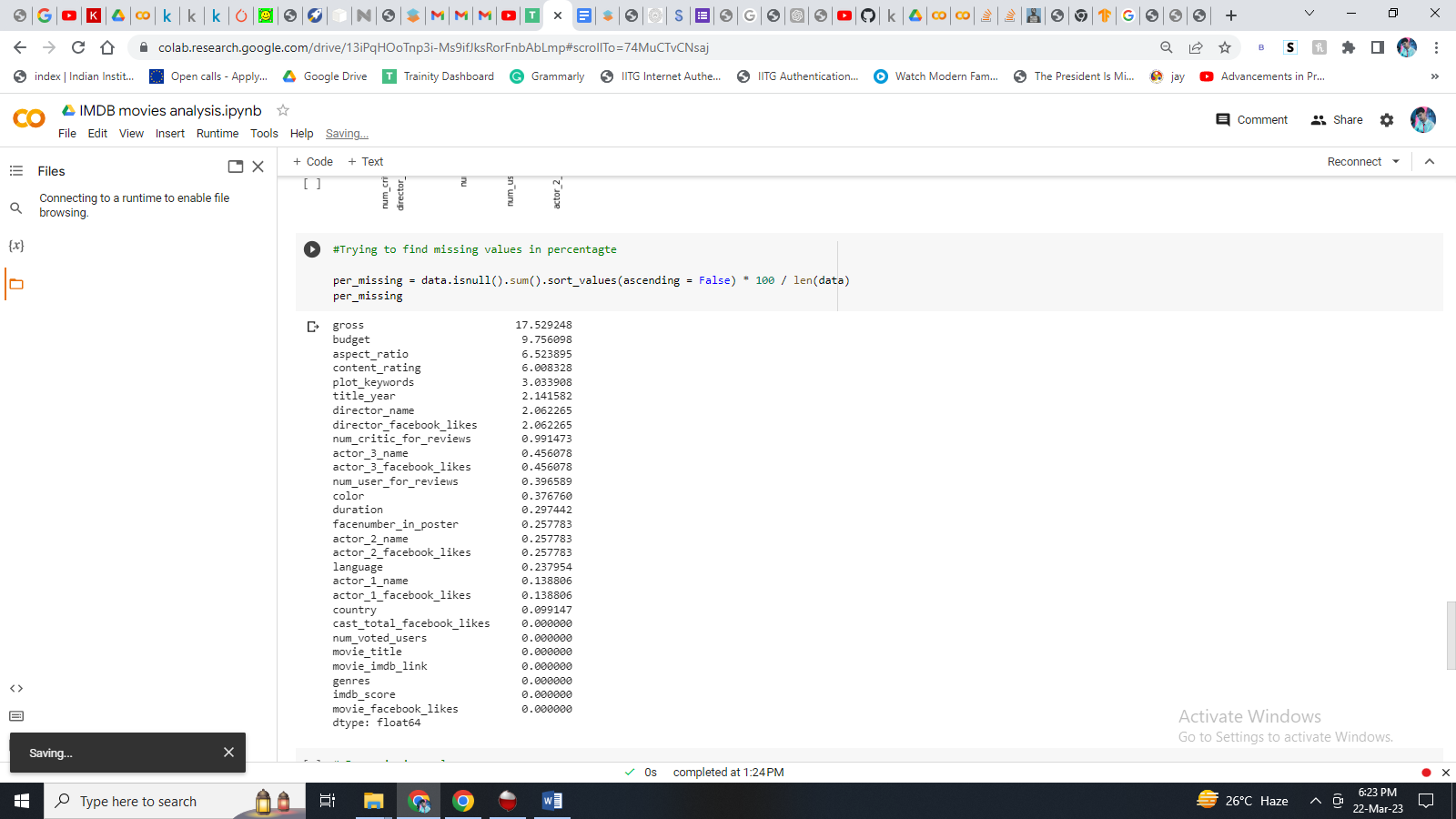


After that we tried to find out the missing values are available or not, and if available we printed it.

Using the command – data.isnull()

Then we tried to establish the columns with missing values with their respective sum.

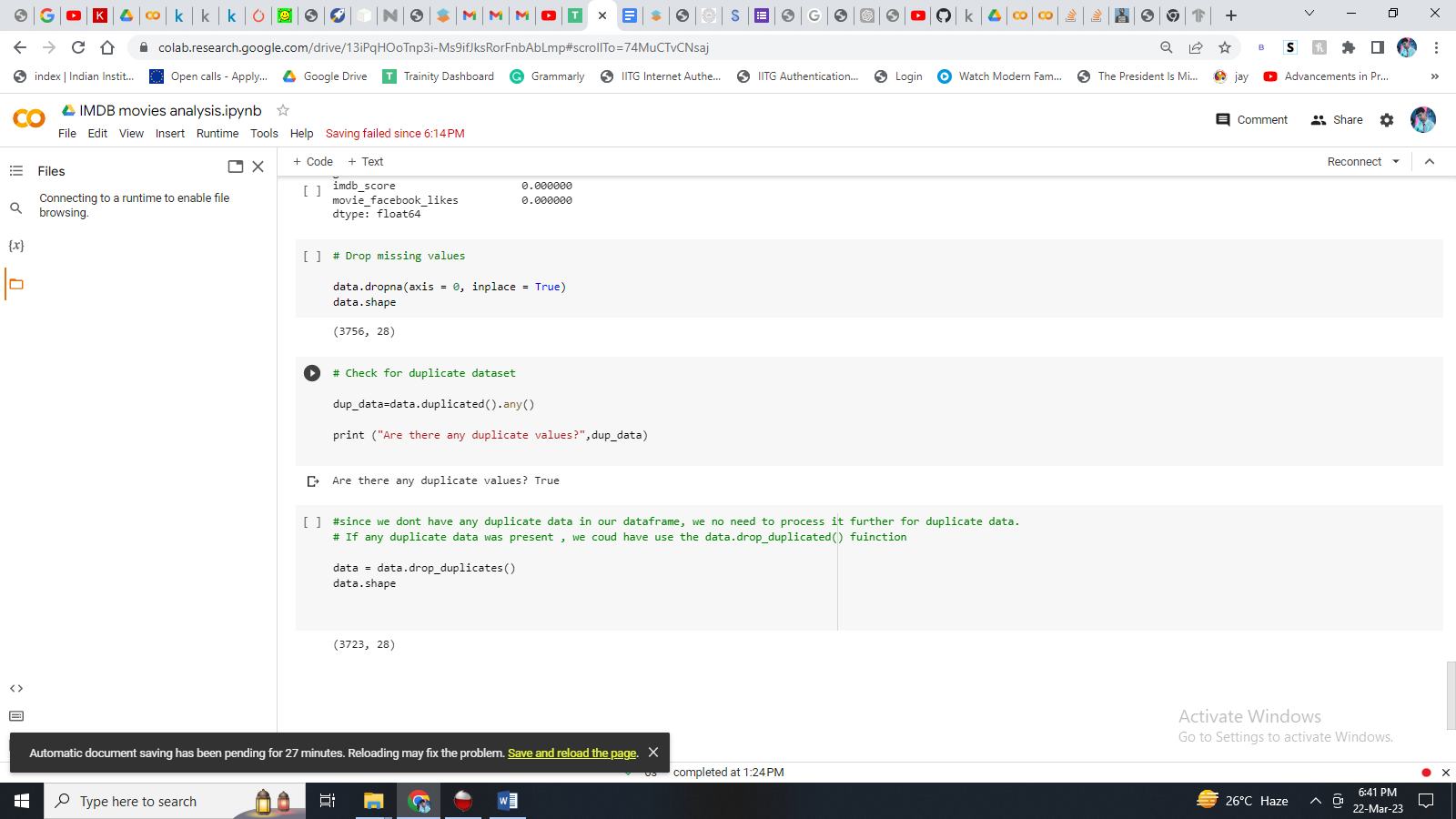




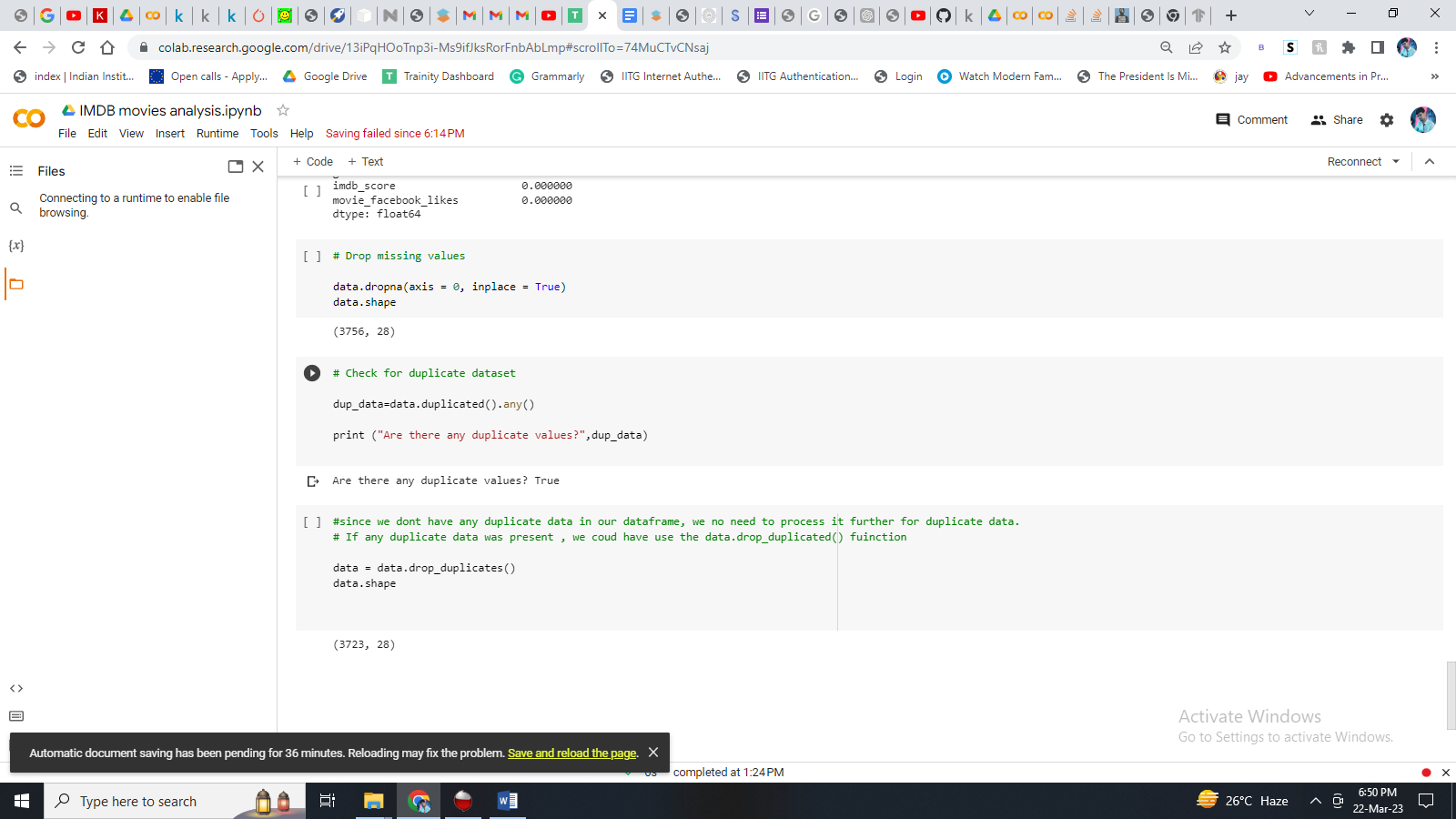
Here, we showed the percentage of missing values in each column.

Gross having highest missing values, followed by budget and aspect\_ratio.

Till now the dataset have **5043 rows × 28** **columns**, altogether including all the missing values, duplicate values and the unnecessary columns not needed for our desired results.

Now we progressed towards removing or dropping the missing values.

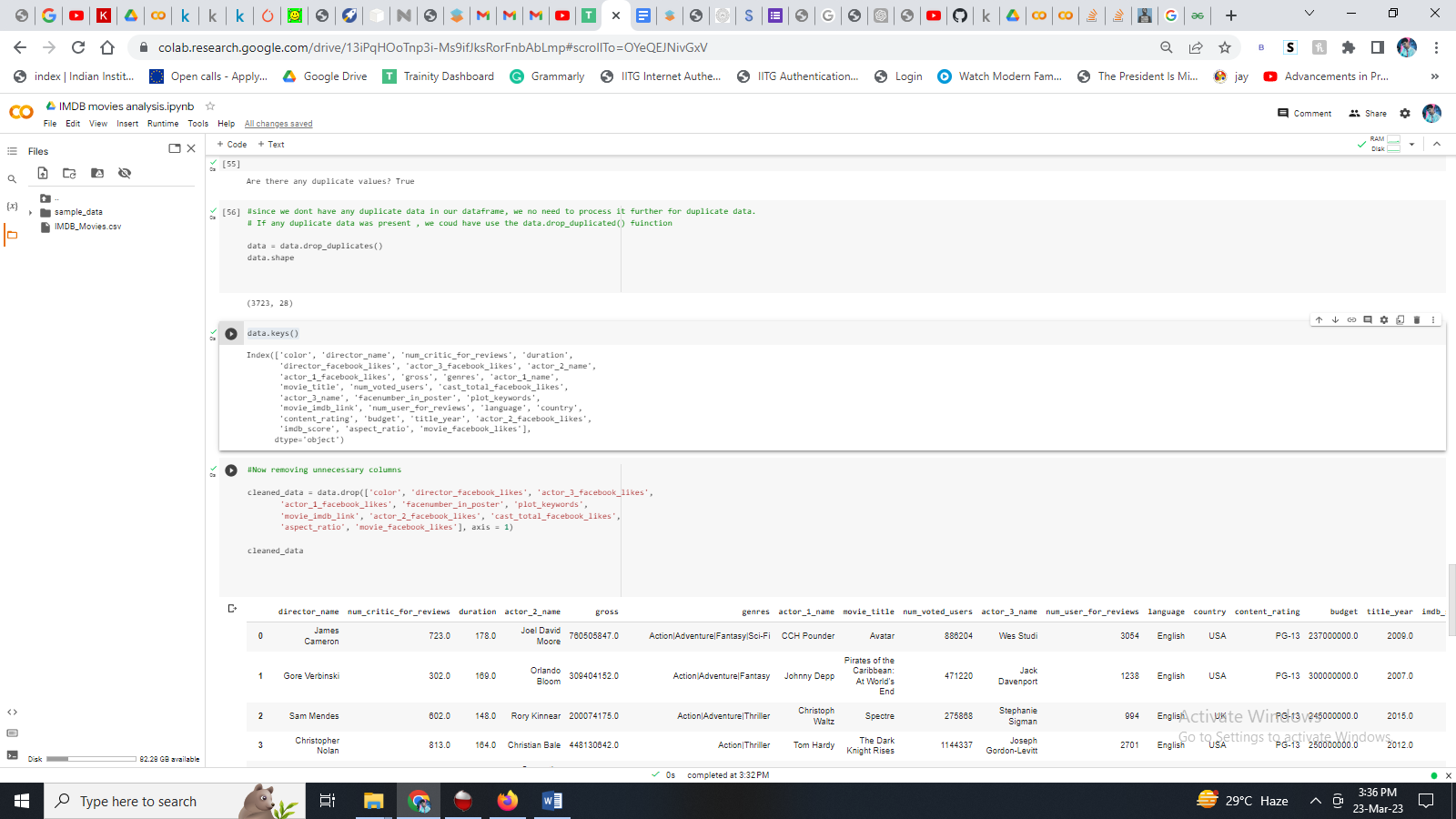
After dropping the missing values finally we are left with **3756 rows \* 28 columns.**

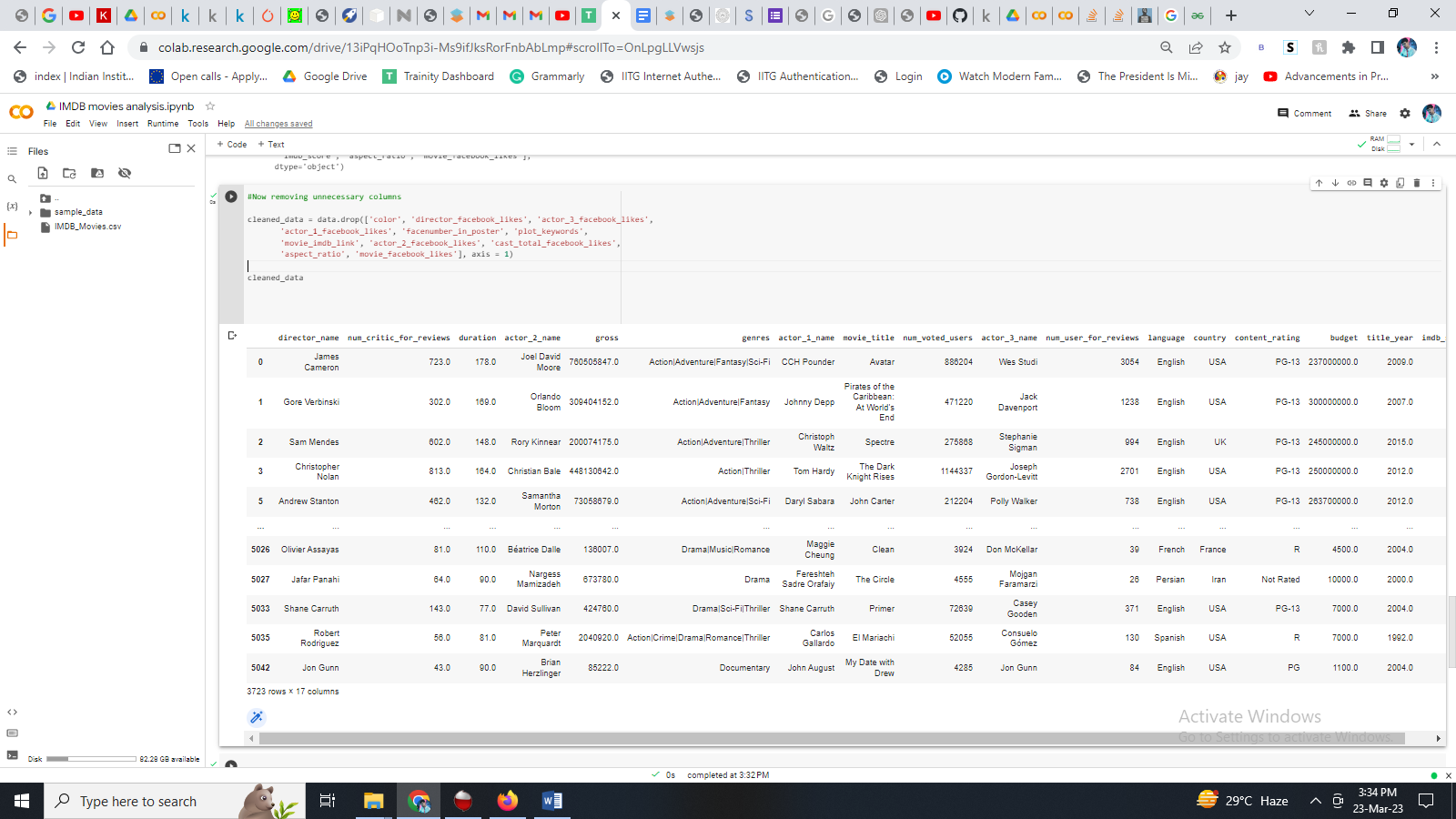
Later we progressed towards identifying is there any duplicate values available, and if available remove from the dataset for cleaning the dataset.

After dropping the duplicate values finally we are left with **3723 rows \* 28 columns.**

Now dropping unnecessary columns, which is not required for our work.

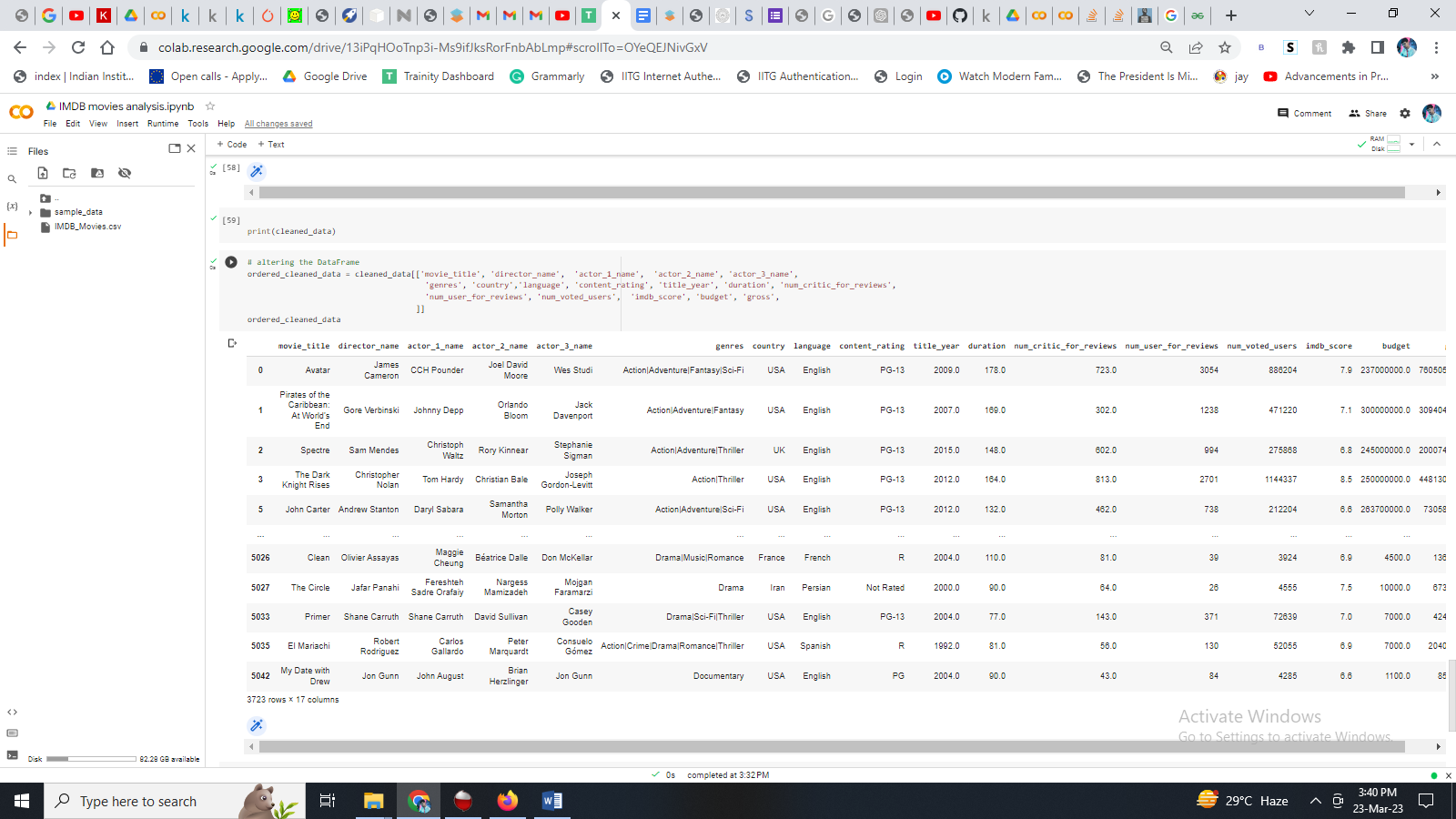
**Initial columns = 28**



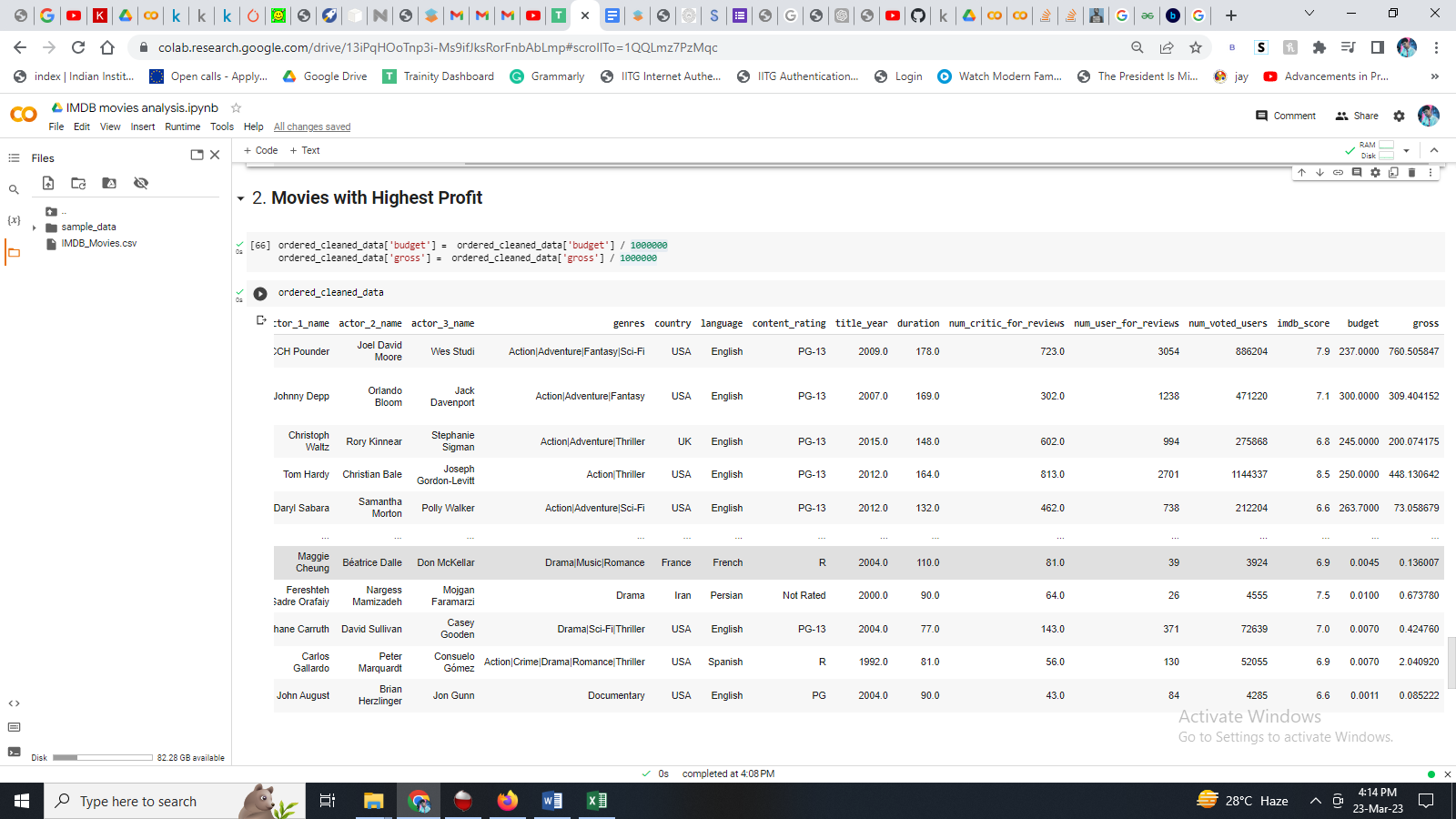
**After deleting unnecessary columns left = 17**

Finally to make the data more readable and usable we re-ordered the columns.

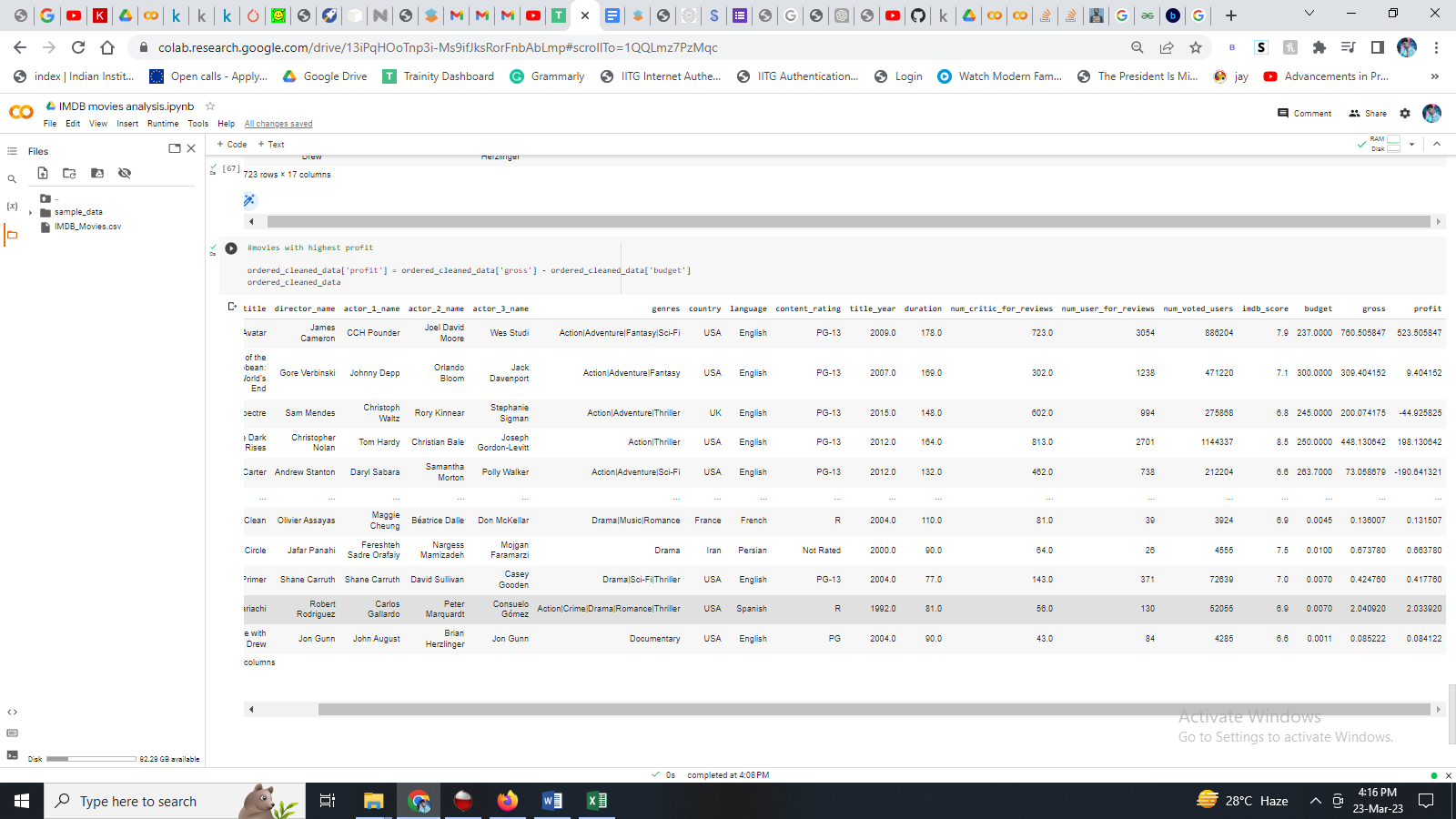
'movie\_title', 'director\_name',  'actor\_1\_name',  'actor\_2\_name', 'actor\_3\_name','genres', 'country','language', 'content\_rating', 'title\_year',  'duration', 'num\_critic\_for\_reviews','num\_user\_for\_reviews',  'num\_voted\_users',  'imdb\_score', 'budget', 'gross',



**B. Movies with highest profit:**

Here, I need to 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.

Here my task is to find the movies with the highest profit.

We found out that “Avatar “ is the highest profit generating movie according to given dataset, with a total profit of **$523.505847 million.**

* 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!

**Here my task is to** Find IMDB Top 250.

|  |
| --- |
| The Shawshank Redemption |
| The Godfather |
| The Dark Knight |
| The Godfather: Part II |
| FargoÂ |
| The Lord of the Rings: The Return of the KingÂ |

1. Here are the list of top 250 IMDB movies with the highest IMDb Rating (corresponding to the column: imdb\_score) – Eg.
2. Also it was made sure that that for all of these movies, the num\_voted\_users is greater than 25,000.
3. And at last a rank column is also added demontratring each movies rank with other details.
4. Also, extraction of all the movies in the IMDb\_Top\_250 column which are not in the English language is performed and stored in a new column named Top\_Foreign\_Lang\_Film. – **Buffy the Vampire Slayer**
5. **Detailed result can be viewed in the excel file question 3 sheet.**



* 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.

Here my task is tofind the best directors.

The top 10 best directors are

**director\_name Mean\_imdb\_scores**

**1.Akira Kurosawa 8.700000**

**2.Charles Chaplin 8.600000**

**3.Tony Kaye 8.600000**

**4.Damien Chazelle 8.500000**

**5.Majid Majidi 8.500000**

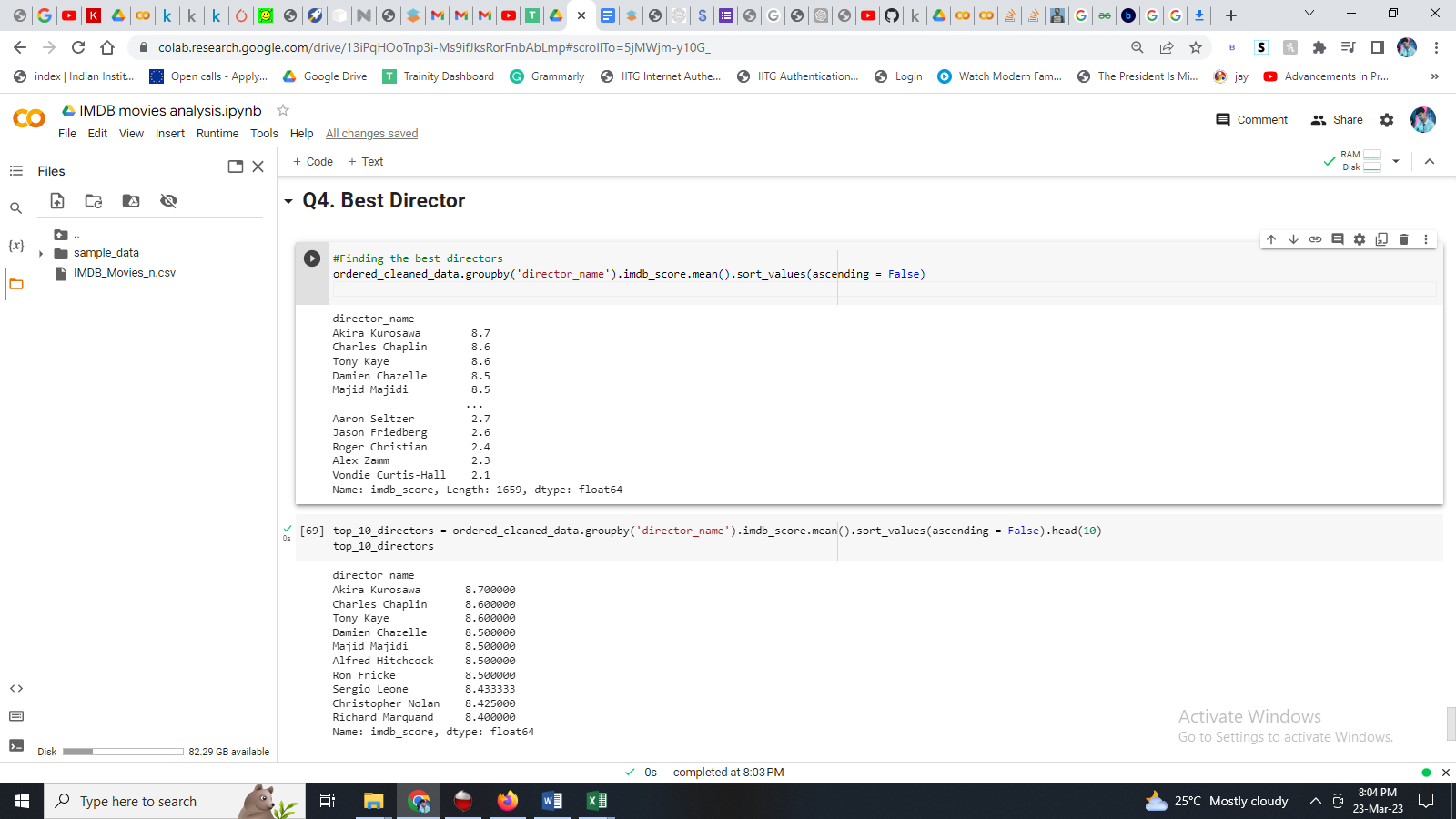
**6.Alfred Hitchcock 8.500000**

**7.Ron Fricke 8.500000**

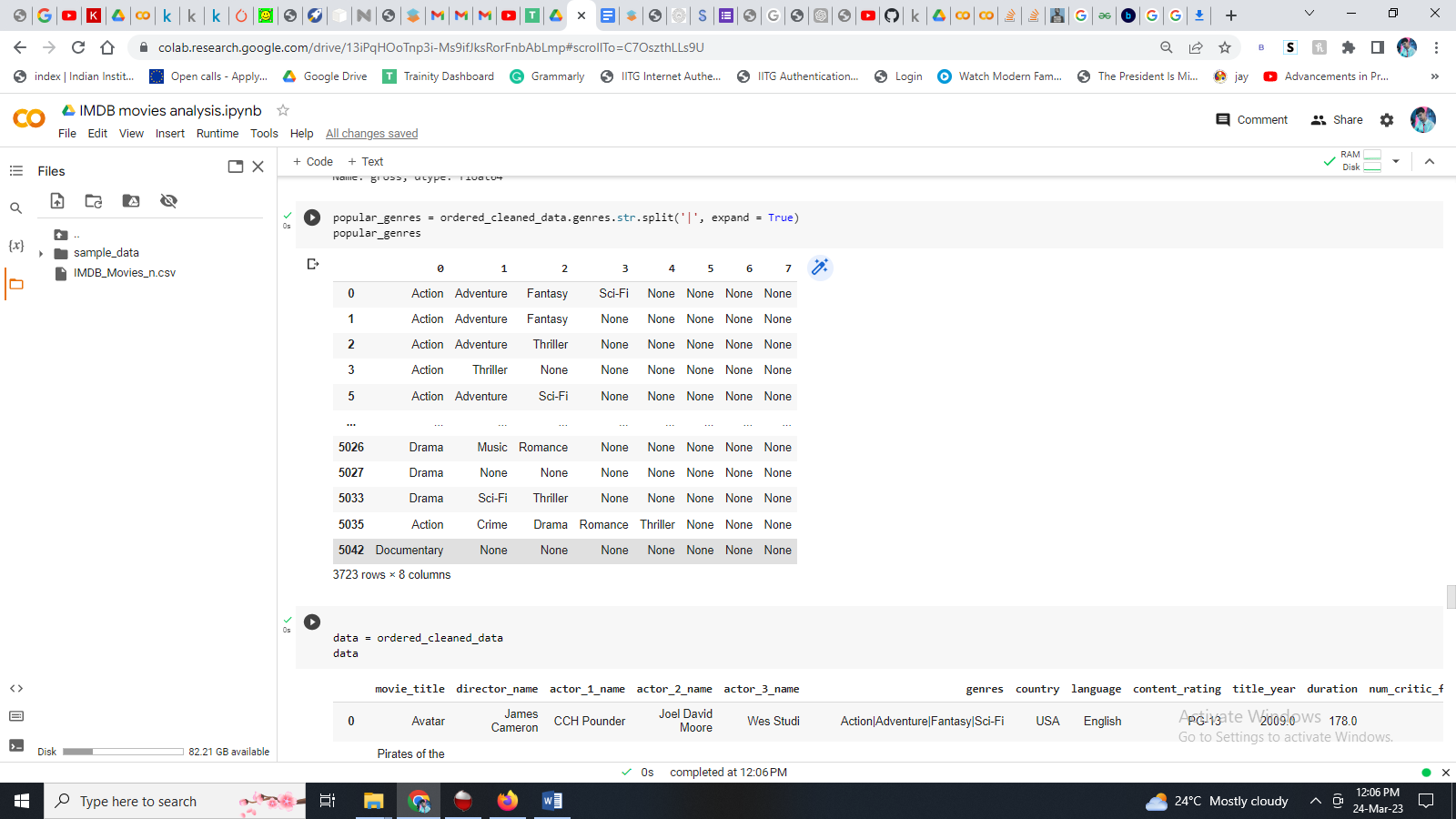
**8.Sergio Leone 8.433333**

**9.Christopher Nolan 8.425000**

**10.Richard Marquand 8.400000**



* 1. **Popular Genres:**

Perform this step using the knowledge gained while performing previous steps.

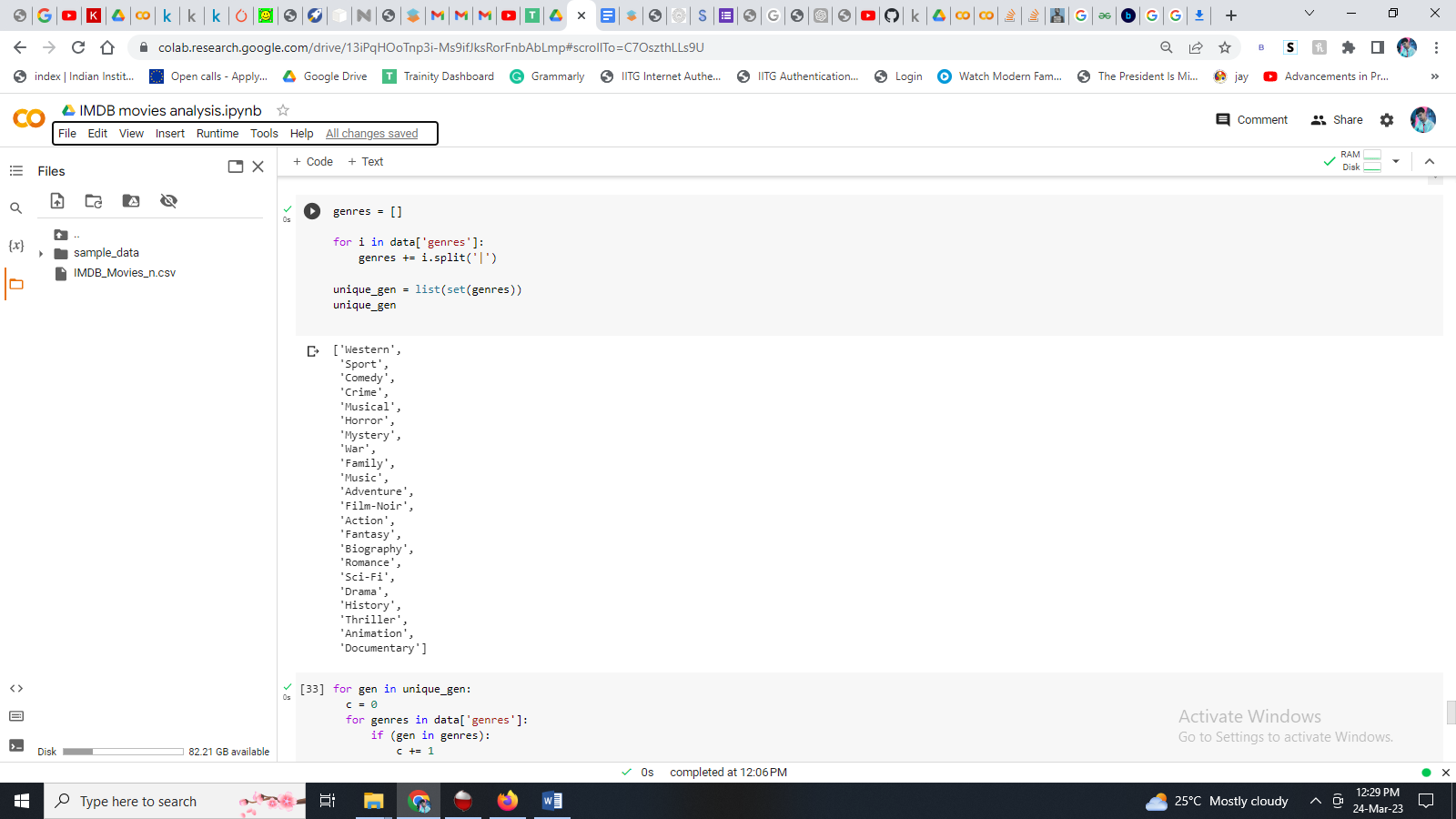
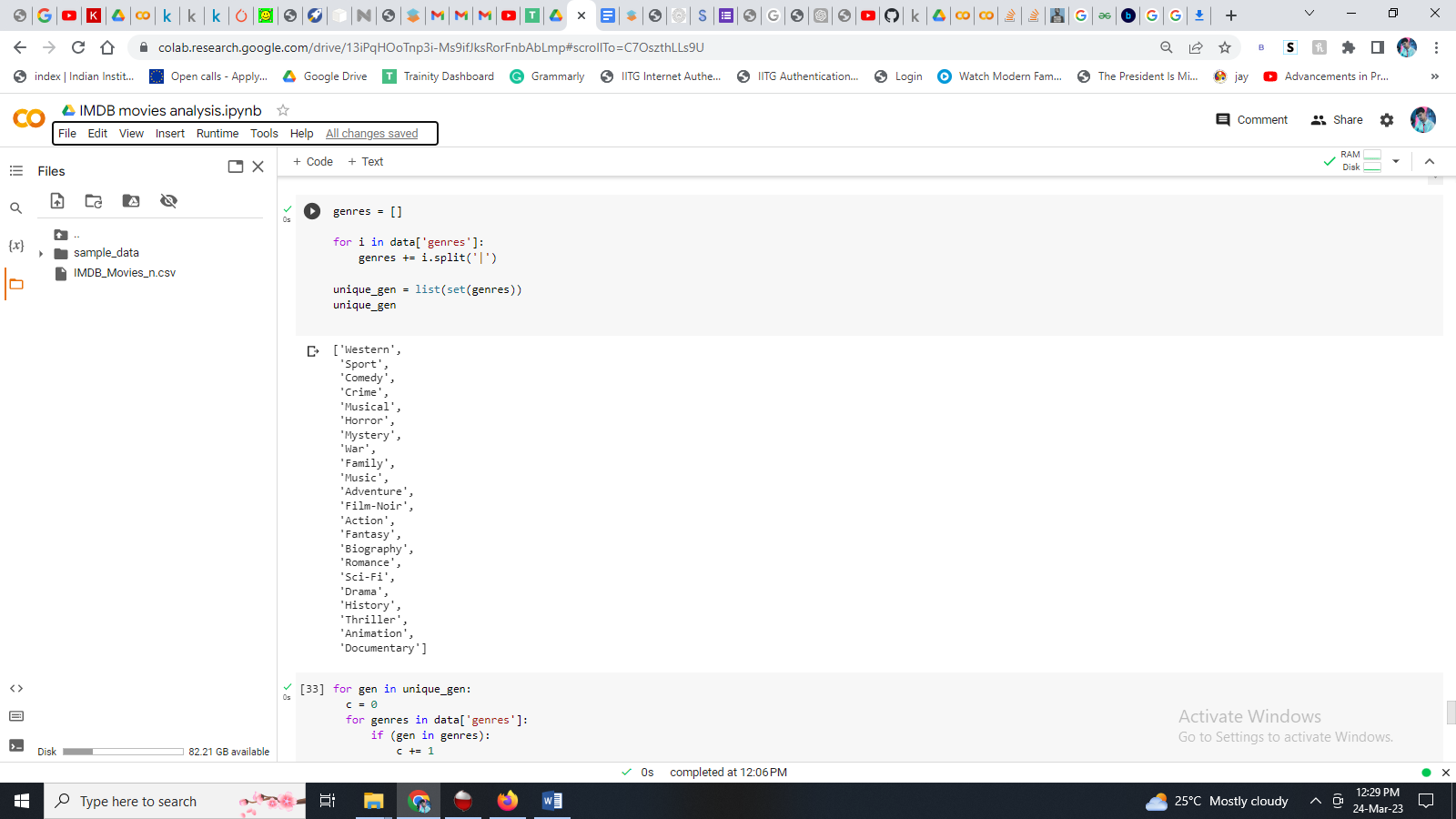
Here our work is to find popular genres.

These are the genres available in the IMDB movies database.

Adventure

Action

Fantasy etc.



Following are the unique geners that are listed in the IMDB mivies database:

Western

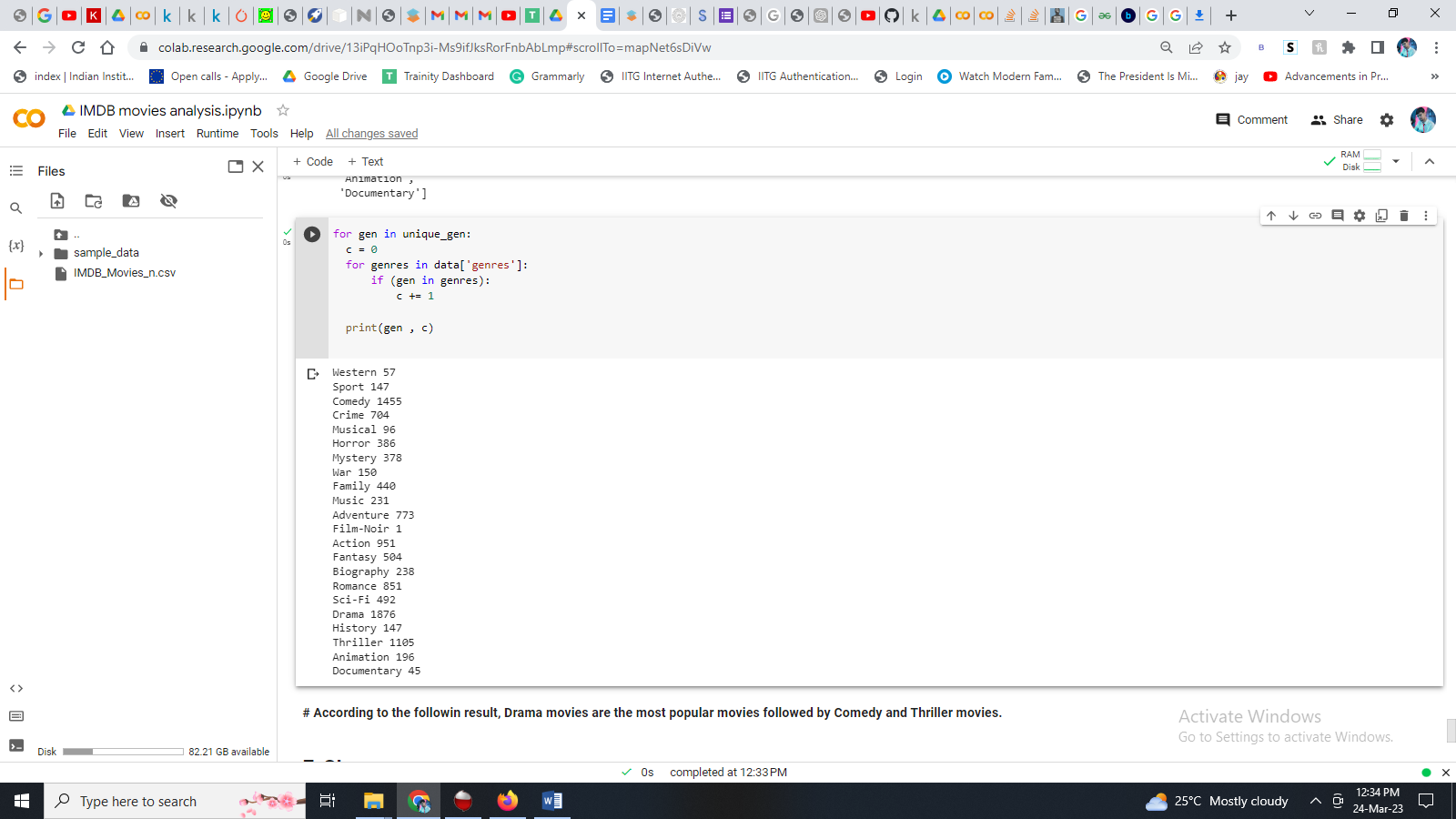
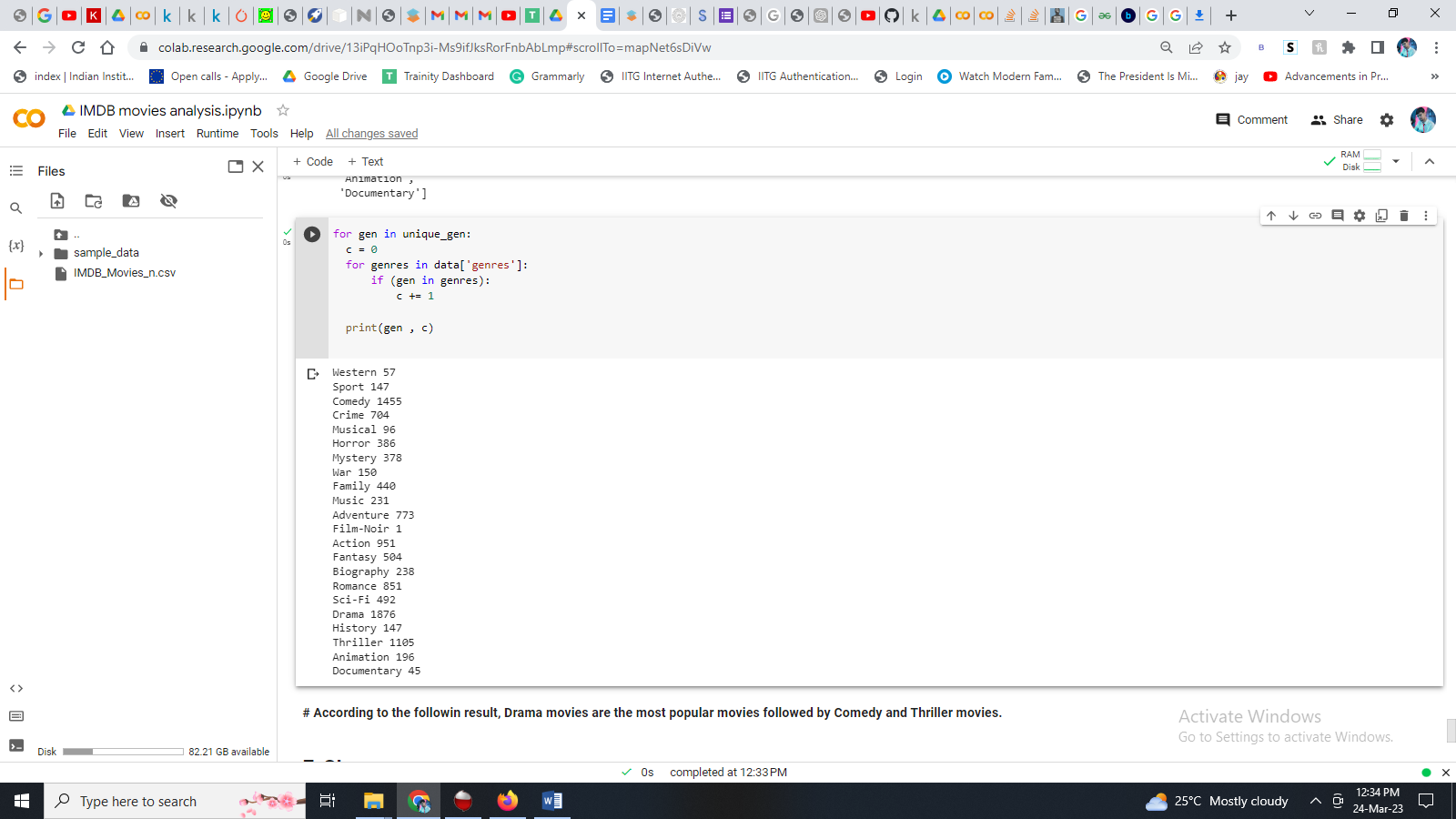
Sport

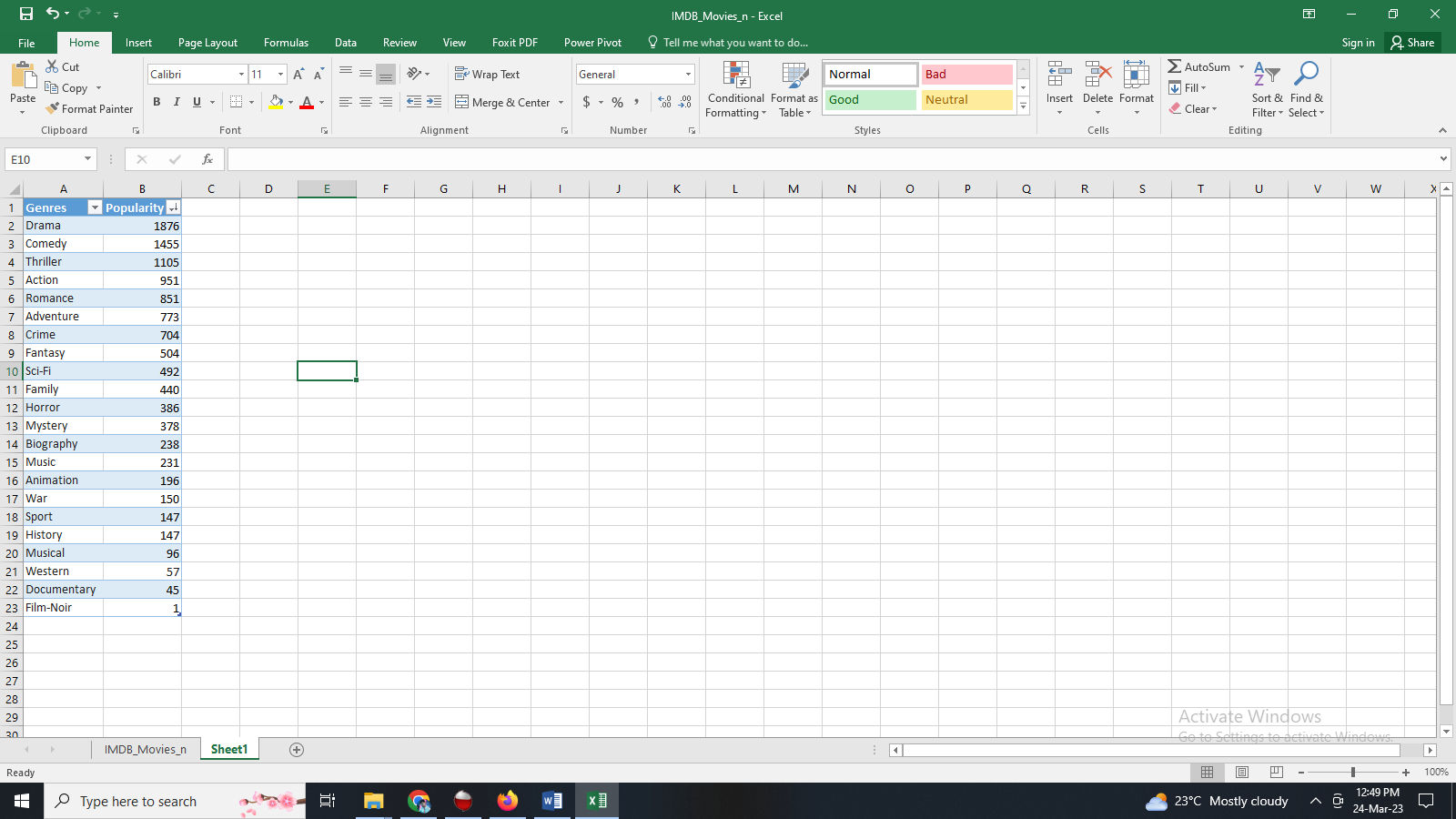
Drama

Action

Adventure etc.

**The following list is the Popularity of the various genres given in the Dataset**



This the final list of unique genres with its popularity based on number of times that genre appeared in the different movies listed on the IMDB dataset given to us.

**The following list is the Popularity of the various genres:**

Drama - 1876

Comedy - 1455

Thriller – 1105

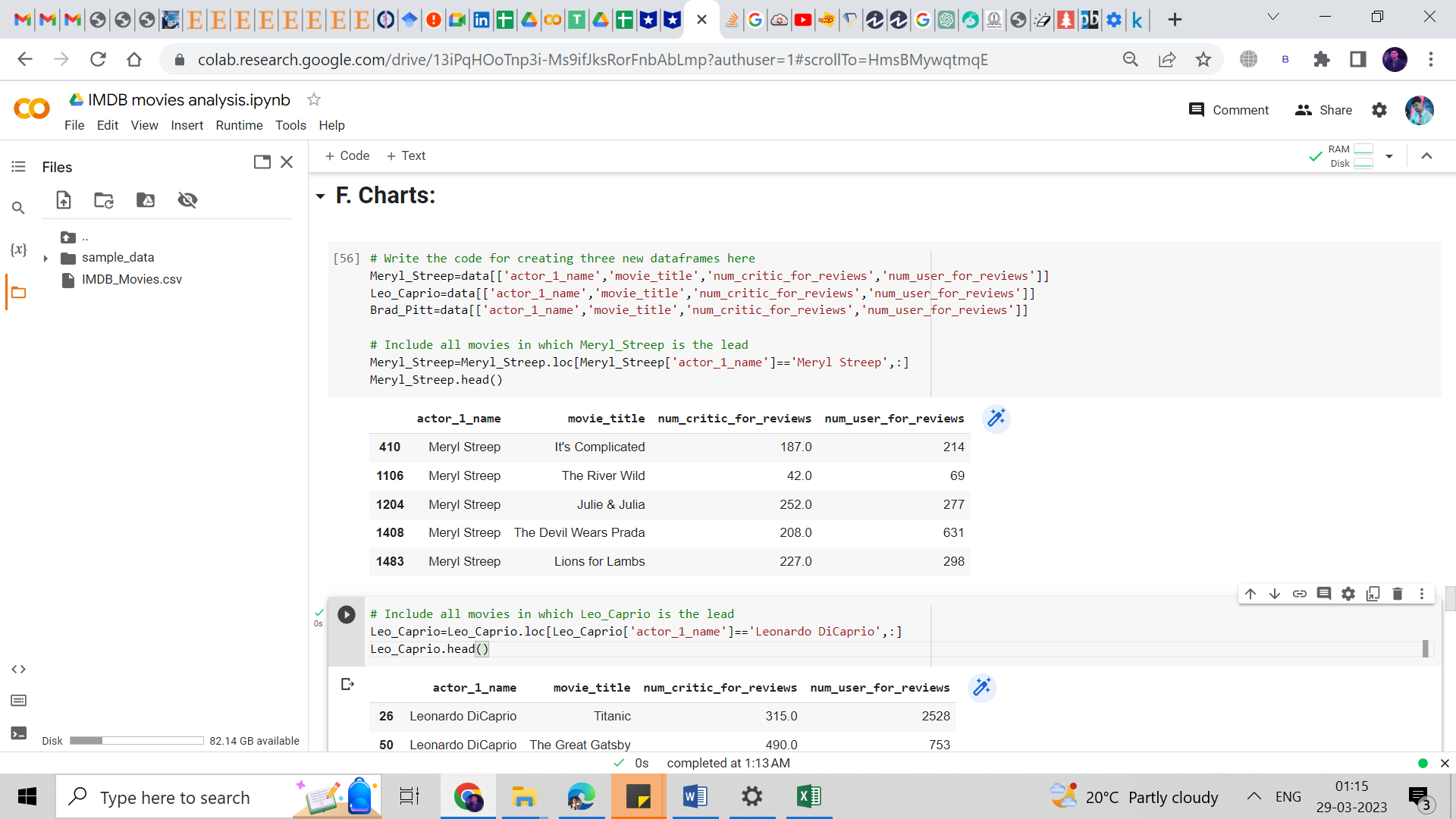
Action - 951

**Conclusion:**

**According to the following result, Drama movies are the most popular movies followed by Comedy and Thriller movies.**

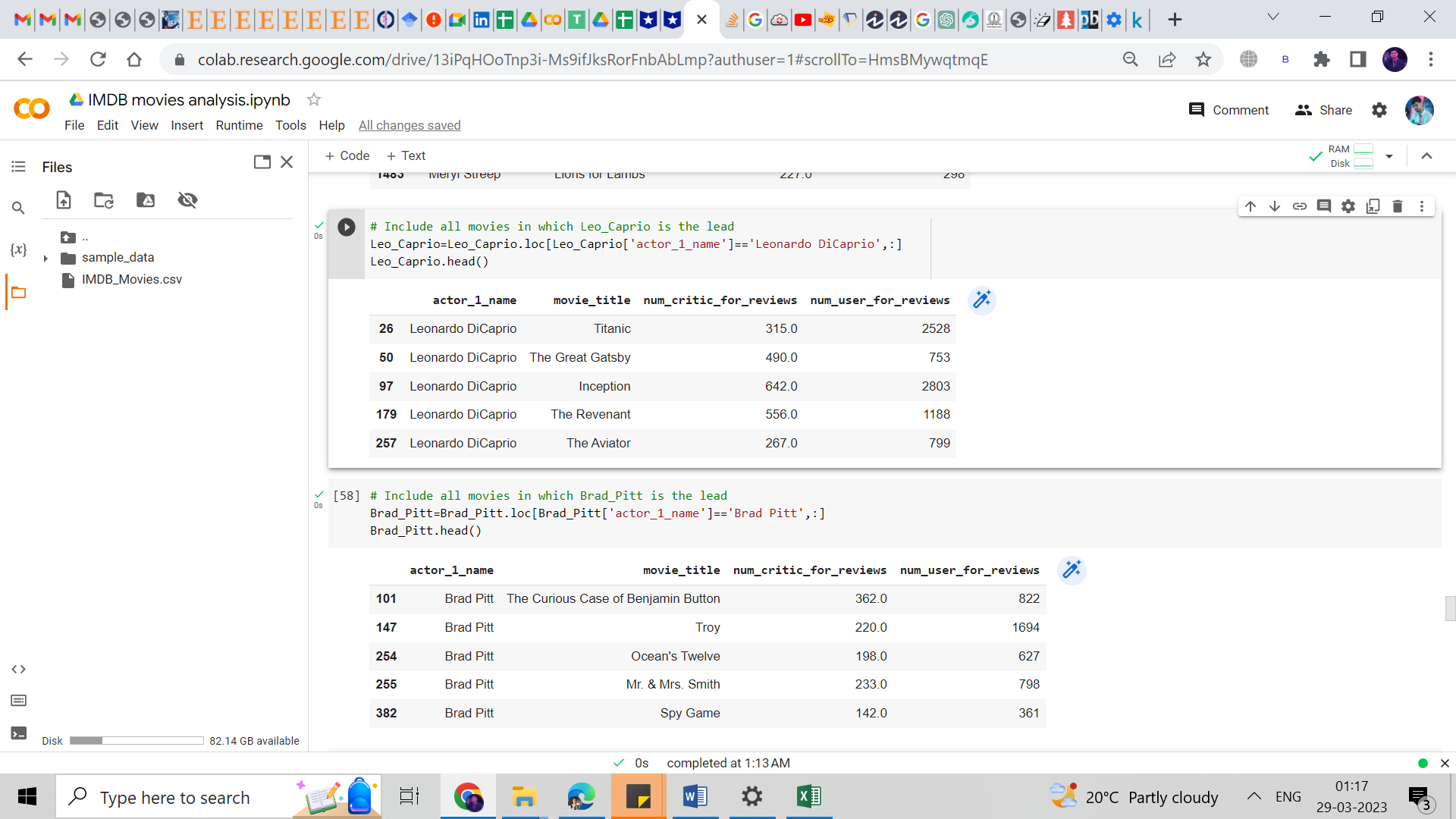
* 1. **Charts:**

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.  
  
  
Here our task is to find the critic-favorite and audience-favorite actors.

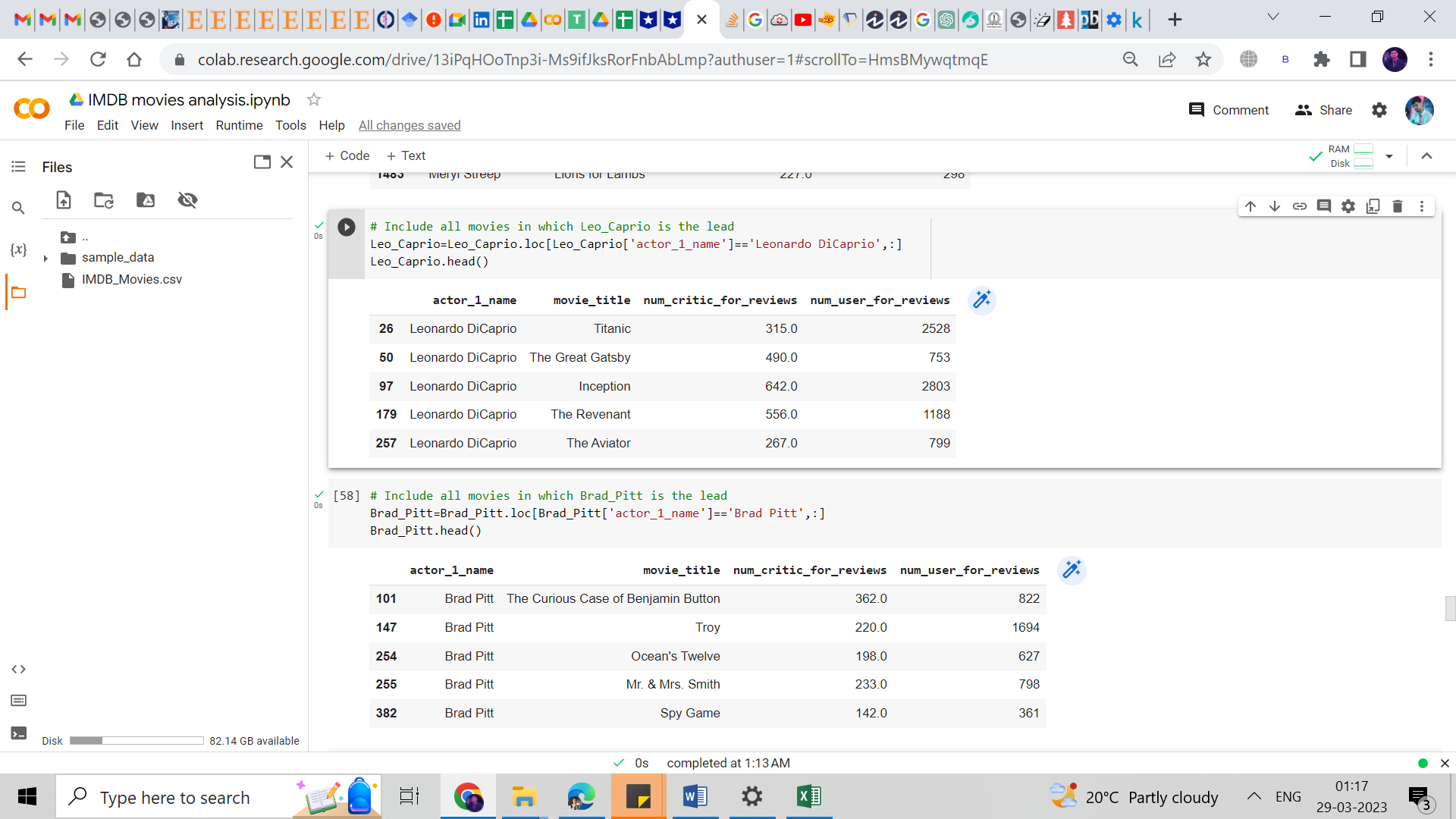
Solution:

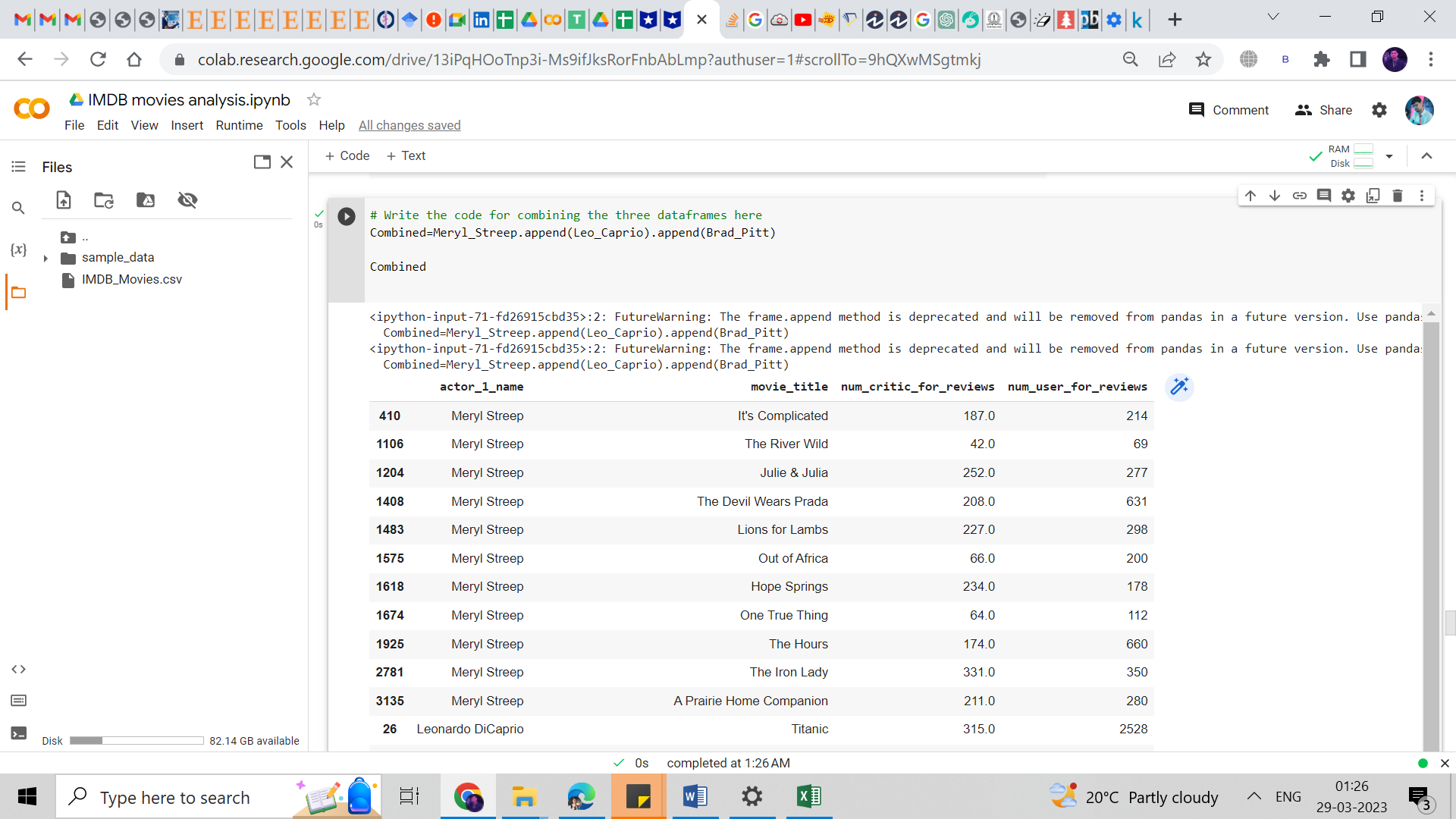
List of, movies where Meryl Steep is the lead actor, along with user and critic reviews.

List of, movies where Leonardo DiCaprio is the lead actor, along with user and critic reviews.

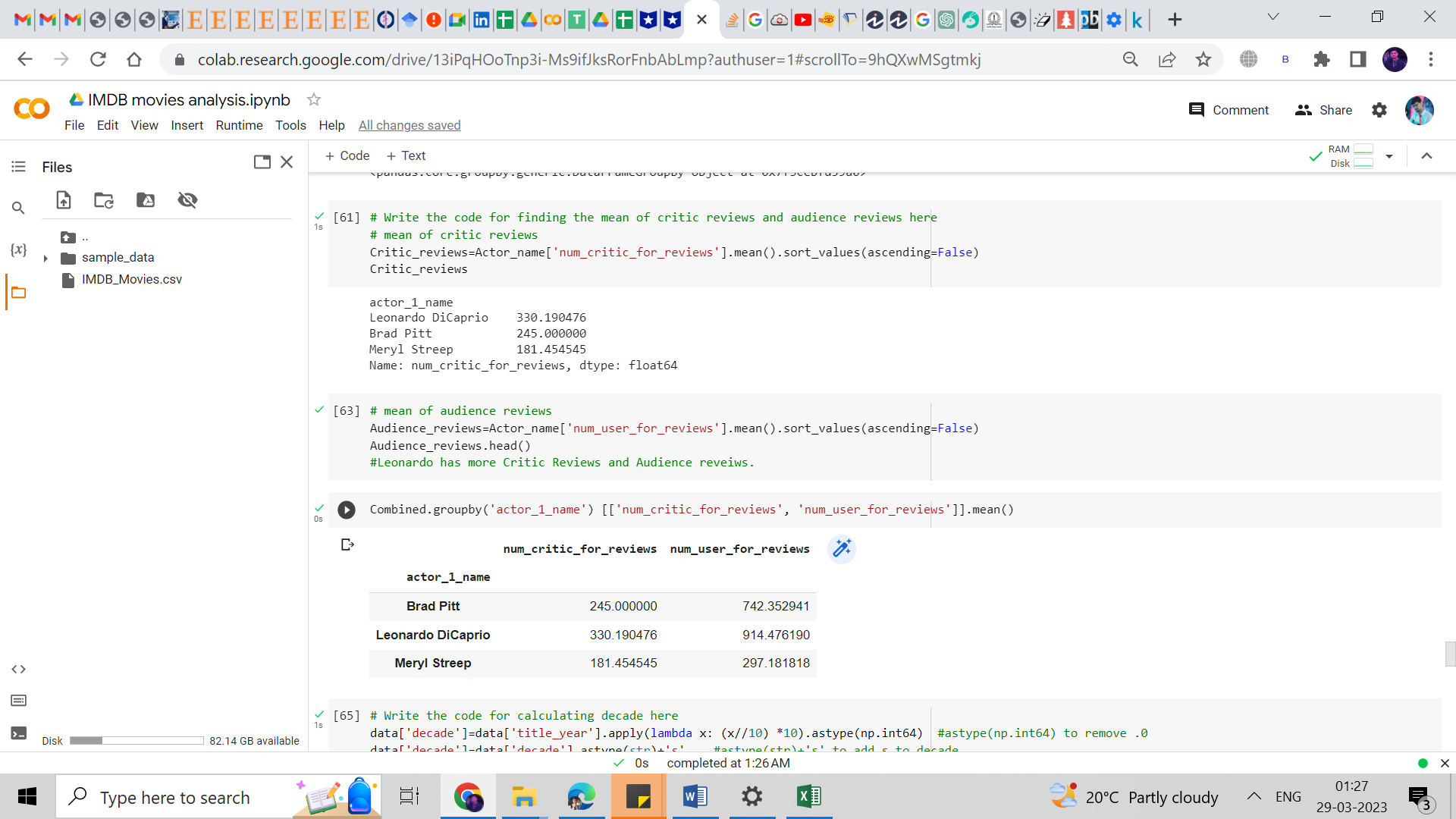


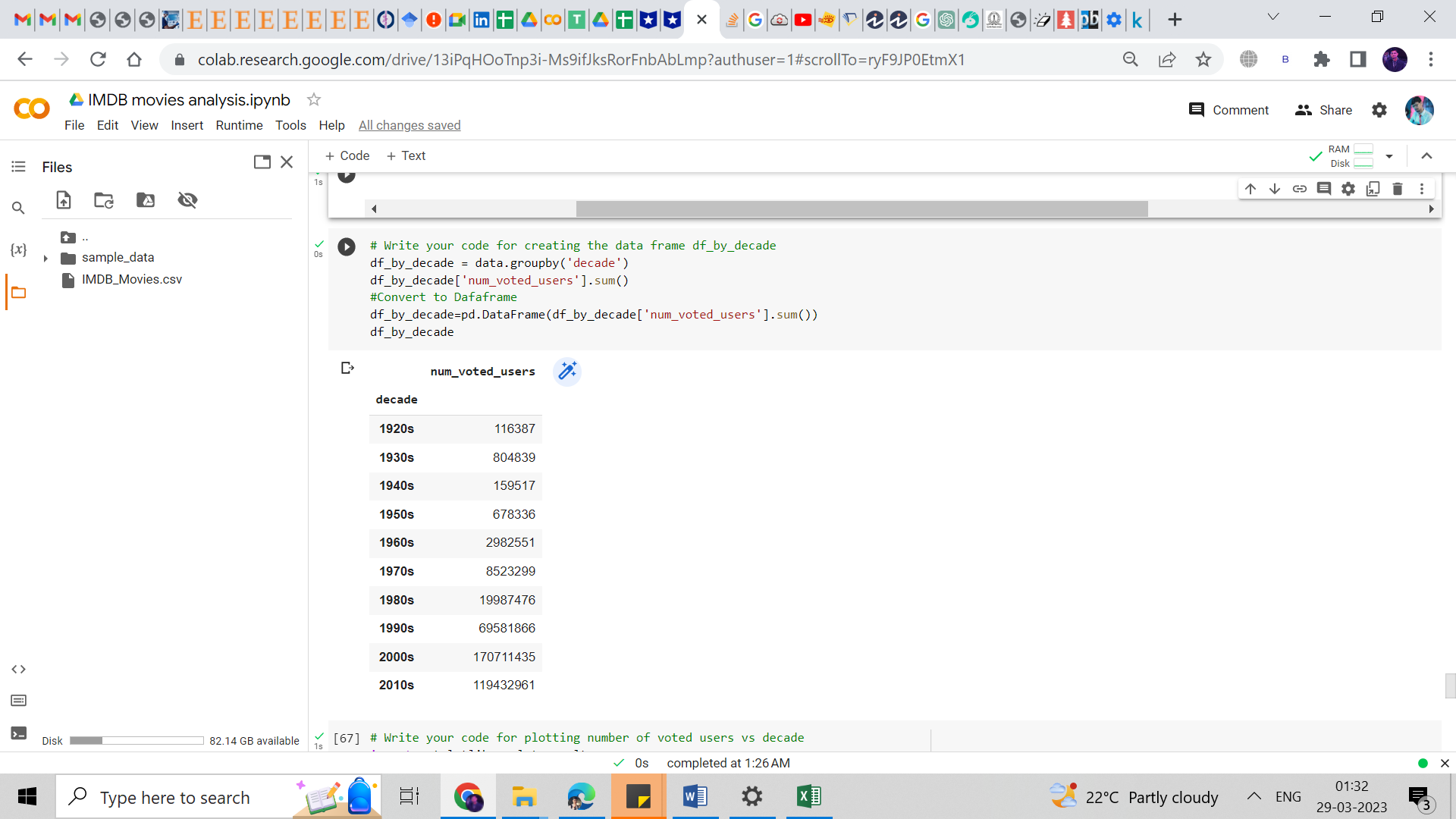
List of, movies where Leonardo DiCaprio is the lead actor, along with user and critic reviews.

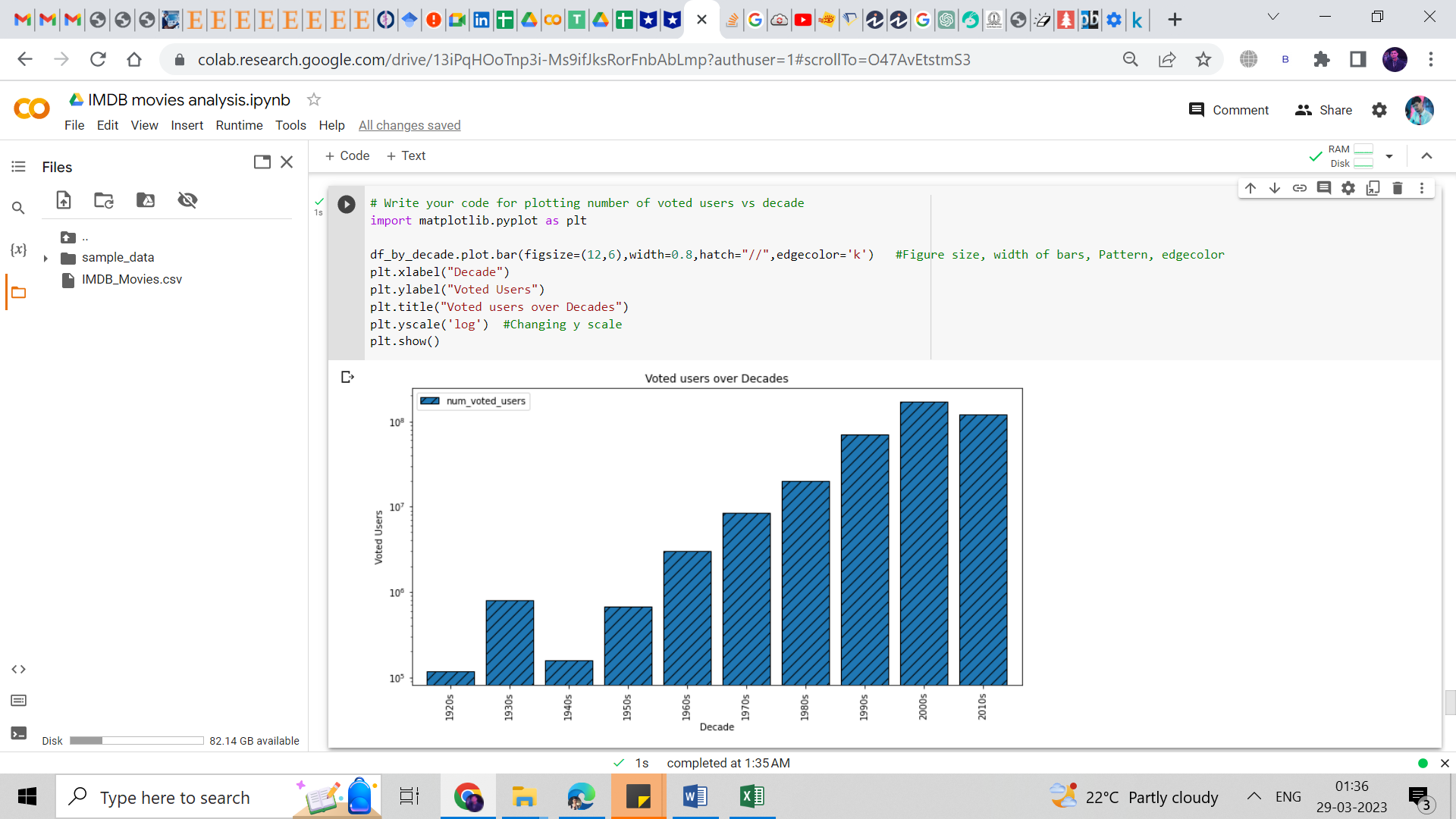


Grouping the combined data frame using the actor\_1\_name column.

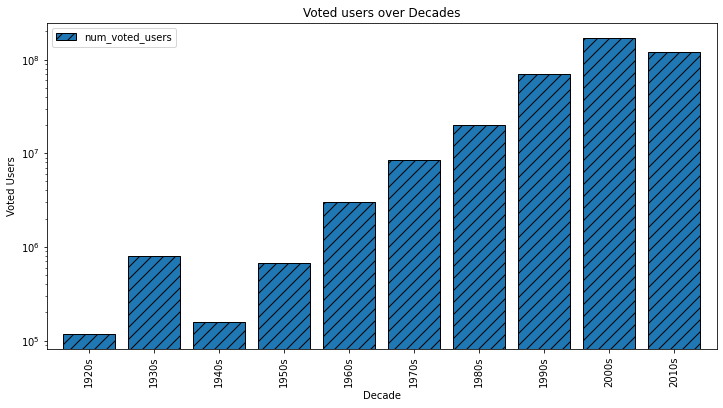
Find the mean of the num\_critic\_for\_reviews and num\_users\_for\_review and identify the actors which have the highest mean.



Sorting the data frame based on the column decade, and grouping it by decade and finding the sum of users voted in each decade.



Observing the change in the number of voted users over decades using a bar chart.



Conclusion:

* Leonardo DiCaprio is the most voted actor by both metrics num\_critic\_for\_reviews and

num\_users\_for\_review, followed by Brad Pitt and Meryl Streep.

* According to the bar chart, the 2010s was the decade when the most number users voted.

**RESULT**

In the making of this report, we used both of our Python and Microsoft Excel knowledge in a real-world example.

In this Project, I achieved Some new things like how to get results from huge amount of data.

**DRIVE LINK**

[**https://drive.google.com/drive/folders/1PixXwW7TEEnwKRBcR5FXddomipsKARzi?usp=share\_link**](https://drive.google.com/drive/folders/1PixXwW7TEEnwKRBcR5FXddomipsKARzi?usp=share_link)

**For a further detailed report, please visit the. ipypnb file in the drive where I have uploaded the file where I built the project using Google collab.**