code.likeagirl.io

Netflix Shows Excel Data Analysis Step-By-Step Case Study!

Uzaifa Memon

8–10 minutes

A Case Study Demonstrating The Data Analysis Process For Beginners Using Microsoft Excel.

Data analysis is a complex and comprehensive process that can frustrate beginners. While Excel initially seems to be a simple tool, that's just a facade. Excel is almost everything a data analyst needs; it is the software data analysts

spend most of their time with. So, if you are looking for an extensive guide, here's one curated especially for level 1 Excel users!

Photo by Alexander Shatov on Unsplash

Good evening, folks!

Can you deny the importance Netflix has in our lives? I know you can't. The streaming giant has become our primary source of entertainment, especially after the pandemic hit the world. Netflix doesn't only symbolize entertainment, though. For geeks of the data

world, Netflix is a major source of bucketloads of data, too. Especially if you are a beginner, just testing the waters in this gigantic universe of data, you'll find this data source superbly easy and engrossing–a combination that will mesmerize you as a data analyst.

In this discussion, let me walk you through the basics of the data analysis process, from listing the objective questions to cleaning the outliers and finally deriving insights from data.

This is a three-step guide for beginners, so download the dataset from the link below for free and get started immediately!

Dataset: https://www.kaggle.com/shivamb/netflix-shows

Objective Questions

1. How Many Titles Have A Runtime Of 100 Minutes?

2. Which Is The Title With The Longest Runtime?
3. Which Is The Title With The Least Runtime?
4. How Many Films Are There With TV-G Ratings?
5. How Many Films And Series Have Been Released On The Same Date?

Though you could derive tens and hundreds of questions, I have addressed these five in this guide for simplification. Through these five questions, we'll be analyzing the data using Excel.

Step 1: Studying The Data

While this isn't a step many are talking about, it is quite essential. You, as a data analyst, need to study your data. It doesn't have to be extensive, just sift through the data, note down the column labels and the content of

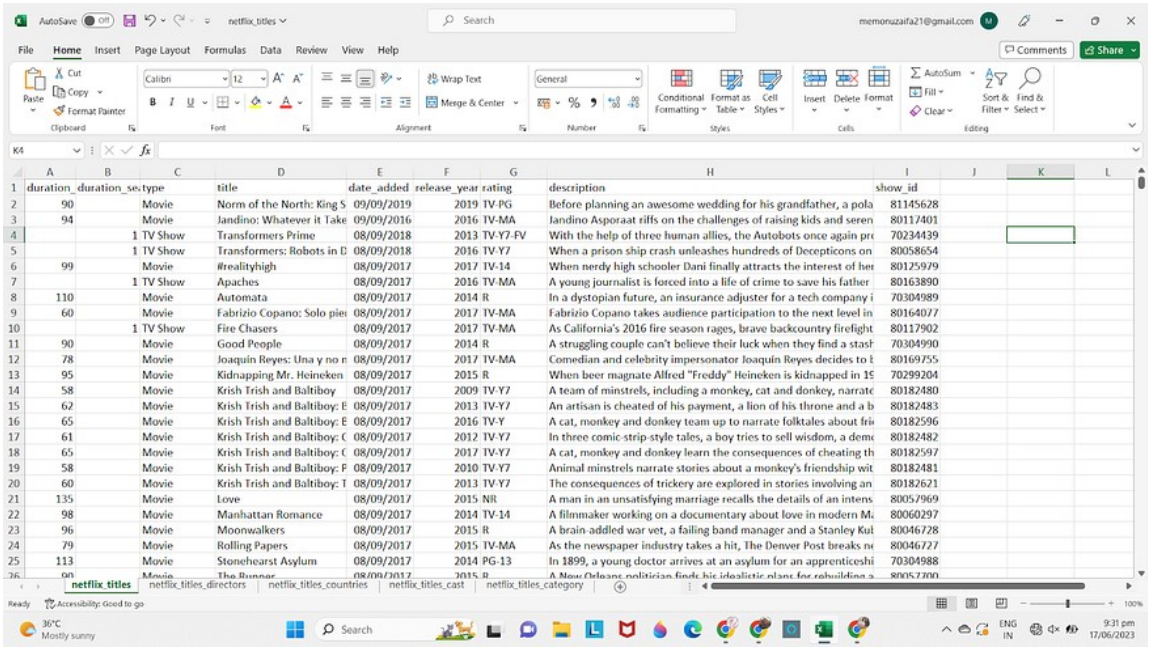
your data before you roll your sleeves.

Here, I've got a *huge amount of data* with thousands of rows and several columns (9 to be precise) describing the title, release date, show id and such details about Netflix's shows. And with the amount of data I have, it will be a lot of work for me–and you, too!

Quick insight:

The number of rows: 6230

The number of columns: 9

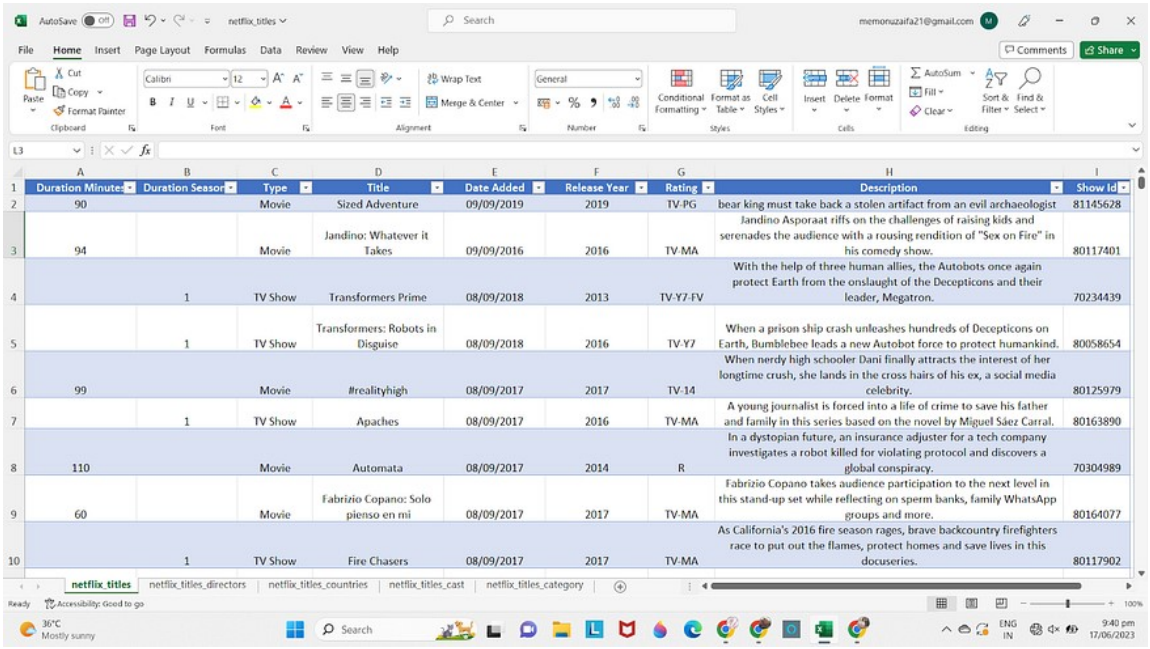


Source: Photo by Author

The first step in your data analysis process

should be converting the data to a table to ease your work. The table provides features normal rows and columns don't, including filters and pivot tables.

After converting the data into a table and applying some basic modifications such as alignment and styling, my table looks!



Duration Minutes	Duration Seasons	Type	Title	Date Added	Release Year	Rating	Description	Show Id
90		Movie	Sized Adventure	09/09/2019	2019	TV-PG	bear king must take back a stolen artifact from an evil archaeologist	81145628
94		Movie	Jandino: Whatever it Takes	09/09/2016	2016	TV-MA	Jandino Asporaat riffs on the challenges of raising kids and serenades the audience with a rousing rendition of "Sex on Fire" in his comedy show.	80117401
	1	TV Show	Transformers Prime	08/09/2018	2013	TV-Y7-FV	With the help of three human allies, the Autobots once again protect Earth from the onslaught of the Decepticons and their leader, Megatron.	70234439
	1	TV Show	Transformers: Robots in Disguise	08/09/2018	2016	TV-Y7	When a prison ship crash unleashes hundreds of Decepticons on Earth, Bumblebee leads a new Autobot force to protect humankind.	80058654
99		Movie	#realityhigh	08/09/2017	2017	TV-14	When nerdy high schooler Dani finally attracts the interest of her longtime crush, she lands in the cross hairs of his ex, a social media celebrity.	80125979
	1	TV Show	Apaches	08/09/2017	2016	TV-MA	A young journalist is forced into a life of crime to save his father and family in this series based on the novel by Miguel Sáez Carral.	80163890
110		Movie	Automata	08/09/2017	2014	R	In a dystopian future, an insurance adjuster for a tech company investigates a robot killed for violating protocol and discovers a global conspiracy.	70304989
60		Movie	Fabrizio Copano: Solo pienso en mí	08/09/2017	2017	TV-MA	Fabrizio Copano takes audience participation to the next level in this stand-up set while reflecting on sperm banks, family WhatsApp groups and more.	80164077
	1	TV Show	Fire Chasers	08/09/2017	2017	TV-MA	As California's 2016 fire season rages, brave backcountry firefighters race to put out the flames, protect homes and save lives in this docuseries.	80117902

Source: Photo by Author

Step 2: Identifying The Outliers And Cleaning The Data

Identify Outliers

Outliers are irregular data patterns that prove

unwanted while analyzing the data. It could be anything that disturbs the structure of your data. For example, if you have all numeric show IDs in one column and find a meaningless text in one row, it is an outlier.

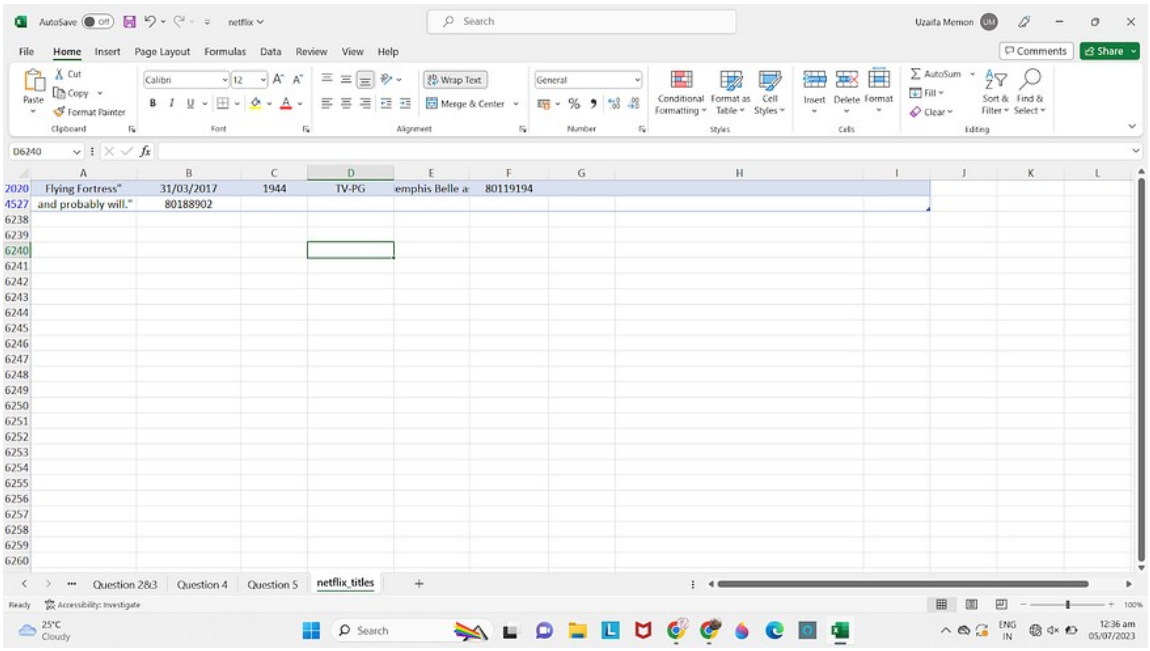
Identifying and handling outliers is the most important step of data analysis. You could do it in several ways, either manually or using a tool. I'm using the Excel table's filter option to help me identify and manage outliers in my dataset.

Outlier 1: Row 2020 and 4527

For row 2020, I'll be just removing the extra dialogue mark.

For row 4527, there are two options available. Either I could delete the row, assuming the data is mistyped and useless, or ask my seniors or data collectors to send me the missing values. But since option 2 isn't

available and I don't find *"and probably will"* as a title for any Netflix show, I'll go with option 1 and delete the row.



Source: Photo by Author

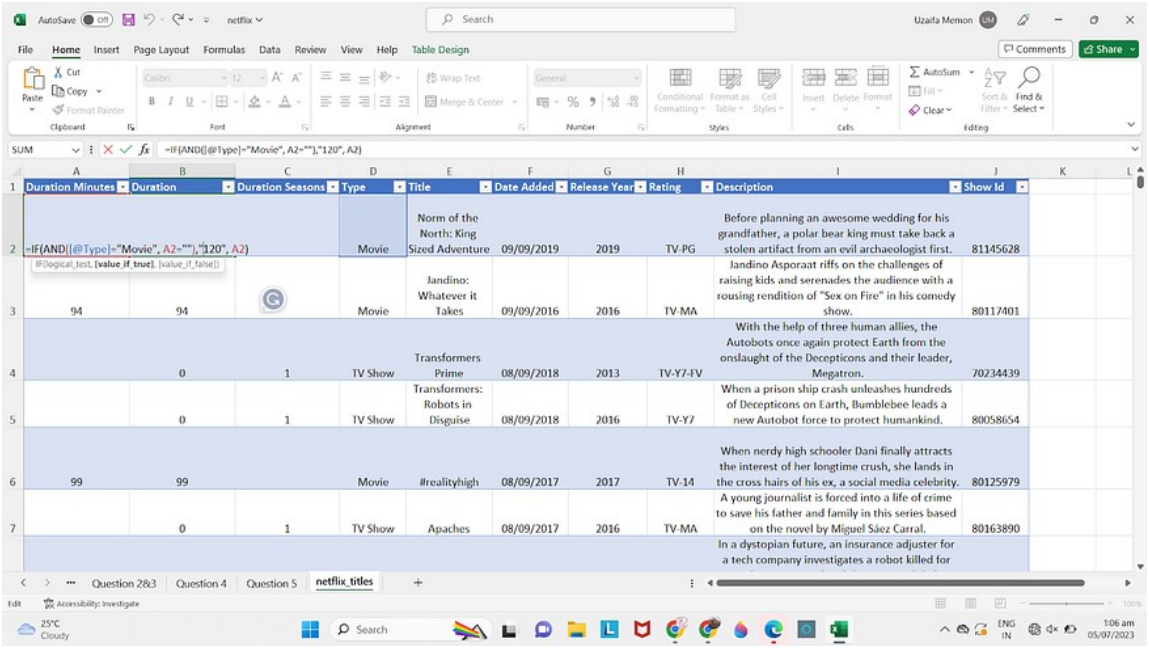
After the outliers were handled, I came across a minor issue. The duration for TV shows is rated in terms of their season count, and there's no missing value in that column. But in the duration column for movies, there are multiple missing values.

So, I have to fill up the data, assuming an average movie timespan is 120 minutes. Here's the formula I'm using to fill up the

data:

IF(AND([@TYPE]==“Movie”, A2=“”), 120, A2)

Quick explanation: If the type mentioned is a movie and the A2 column currently contains the movie's duration is blank, fill up the cell with the value “120,” or else the value should be the one mentioned in the corresponding A2 cell.



Source: Photo by Author

Continue this process for every outlier you come across in the dataset. In my case, I have come across two more in which the date has

been missing. I filled the cell with the date (hypothetical) and moved on.

Identify And Eliminate Duplicate Data

Now that we're done rectifying the outliers, we should consider duplicate data. In this dataset, there are several duplicate values but not the ones that matter, aka the series title.

Step 3: Generating Insights From The Cleansed Dataset

Here's when the interesting part arrives!

Remember the objective questions I listed above? We'll try finding the answers to each of them now.

Answer 1

Our first task is to count the number of titles with a runtime of 100 minutes. To accomplish this, I'll be using the countifs formula. The first task we'll be doing is to create a pivot

table in a new sheet.

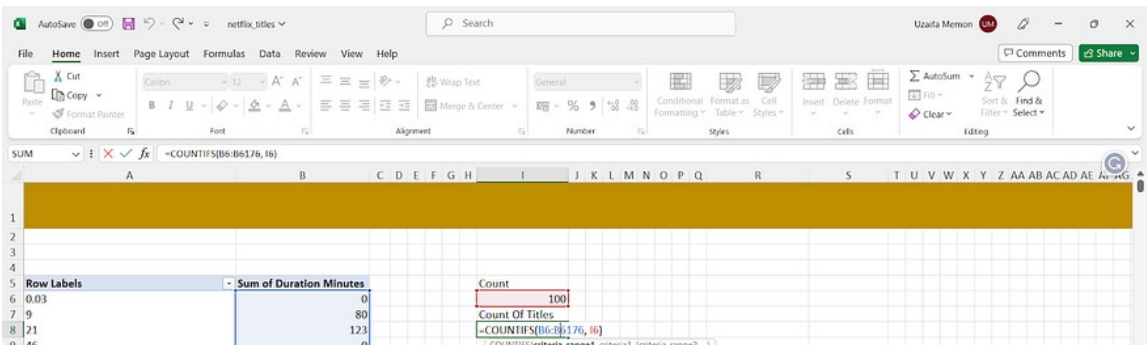
Select the Pivot Table option from the Insert ribbon, and Excel will automatically direct you to a new sheet. Select and drag the Title option to the row field and Duration Minutes to the value field. And tada, here's your pivot table ready!

The formula using which I am counting the number of titles present in the range B6: B6176 with the duration mentioned in the R6 cell is:

COUNTIFS(B6:B6176, R6)

A cell reference (R6) is used to add a dynamic touch. If the number in the R6 cell changes, so will the count.

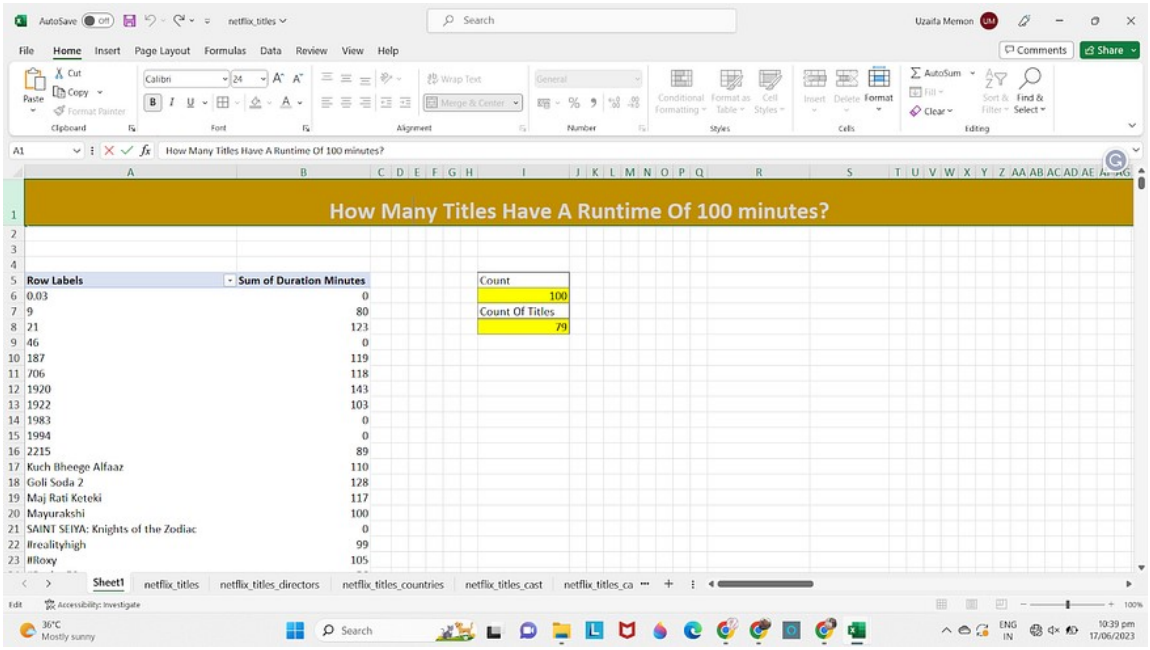
One formula for all the durations!



10	187	119
11	706	118
12	1920	143
13	1922	103
14	1983	0
15	1994	0
16	2215	89
17	Kuch Bheege Alfaaz	110
18	Goli Soda 2	128
19	Maj Rati Keteki	117
20	Mayurakshi	100
21	SAINT SEIYA: Knights of the Zodiac	0
22	#realityhigh	99
23	#Roxy	105

Source: Photo by Author

Output:



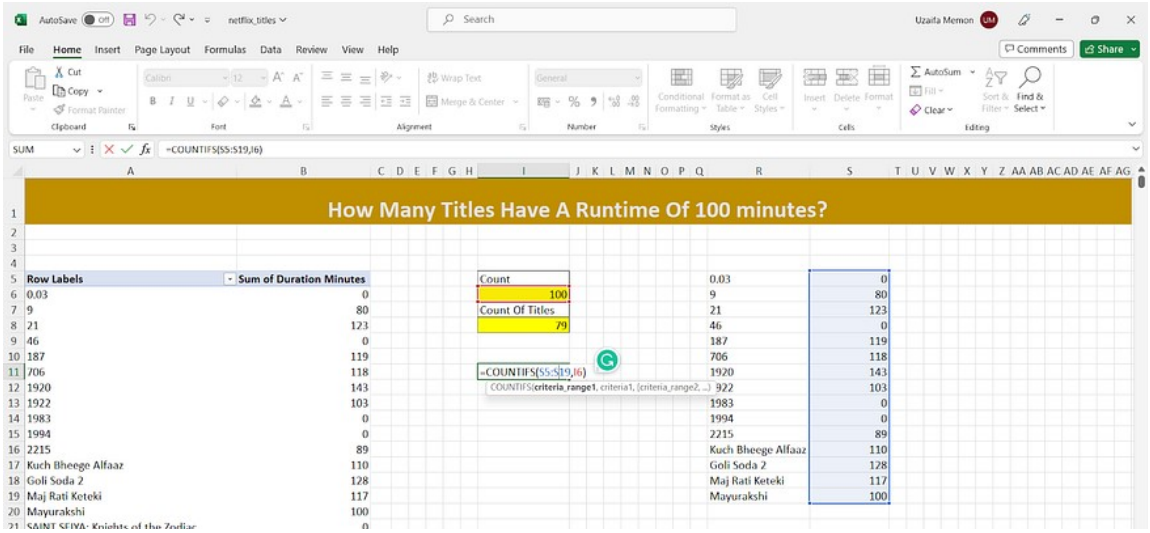
How Many Titles Have A Runtime Of 100 minutes?

Row Labels	Sum of Duration Minutes
0.03	0
9	80
21	123
46	0
187	119
706	118
1920	143
1922	103
1983	0
1994	0
2215	89
Kuch Bheege Alfaaz	110
Goli Soda 2	128
Maj Rati Keteki	117
Mayurakshi	100
SAINT SEIYA: Knights of the Zodiac	0
#realityhigh	99
#Roxy	105

Count
100
Count Of Titles
79

Source: Photo by Author

Let us crosscheck our formula and the results using a subset of our dataset quickly!



How Many Titles Have A Runtime Of 100 minutes?

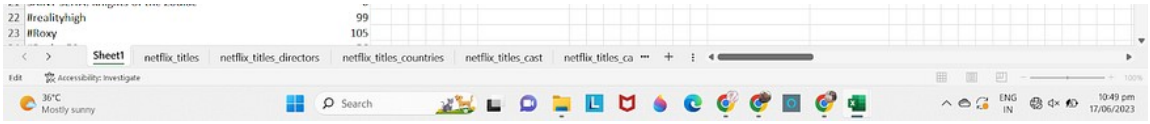
Row Labels	Sum of Duration Minutes
0.03	0
9	80
21	123
46	0
187	119
706	118
1920	143
1922	103
1983	0
1994	0
2215	89
Kuch Bheege Alfaaz	110
Goli Soda 2	128
Maj Rati Keteki	117
Mayurakshi	100

Count
100
Count Of Titles
79

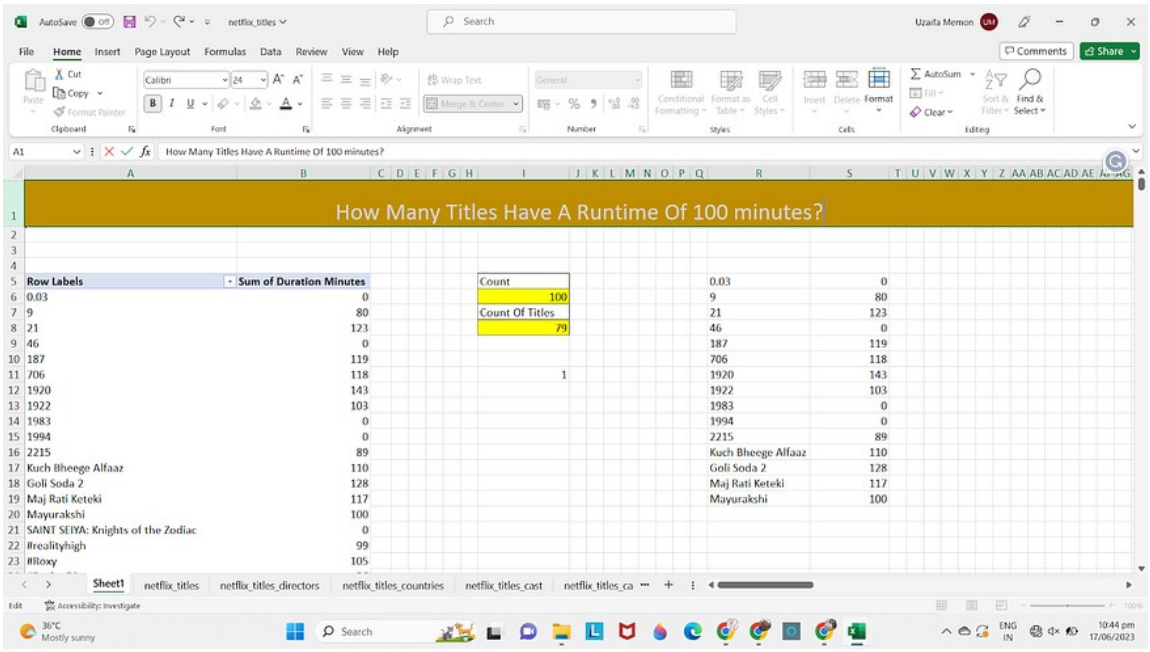
=COUNTIFS(S5:S19,I6)

COUNTIFS(criteria_range1, criteria1, [criteria_range2, ...])

0.03	0
9	80
21	123
46	0
187	119
706	118
1920	143
922	103
1983	0
1994	0
2215	89
Kuch Bheege Alfaaz	110
Goli Soda 2	128
Maj Rati Keteki	117
Mayurakshi	100



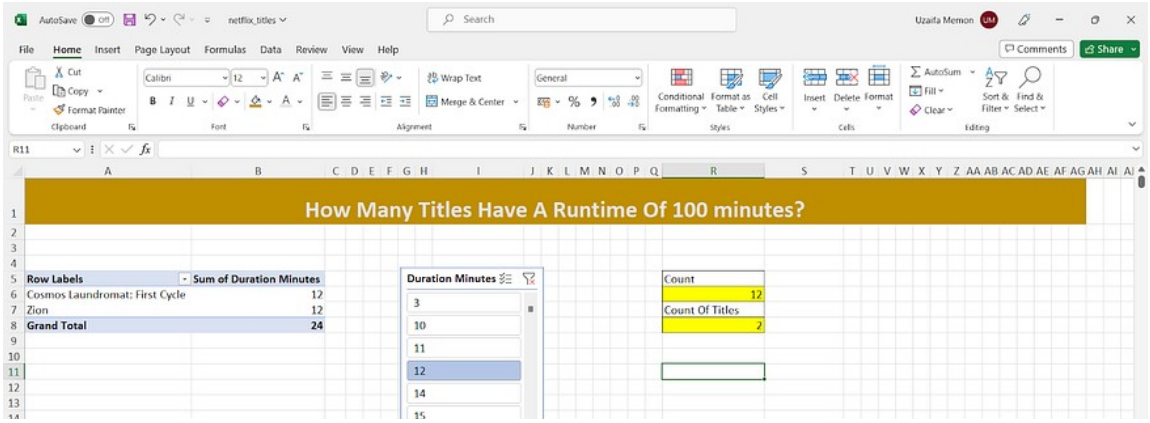
Source: Photo by [Author]

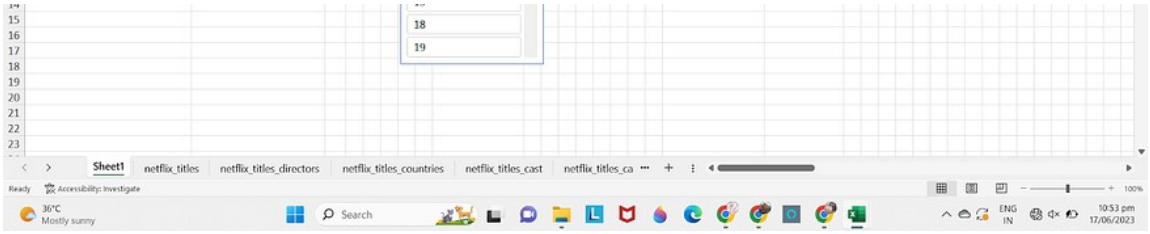


Source: Photo by [Author]

Now, how about a shortcut?

You can easily add a slicer to the pivot table, and it'll list the available durations and list only those values matching the criteria. Slicer acts as a filter and can easily simplify your work.





Source: Photo by Author

As you can see above, the slicer returns the same count as the formula I wrote, which means our first question has an answer!

Answer 2, 3:

Which Is The Title With The Longest Runtime?

The answer to this question is quite simple, but I'm breaking it down into three parts to simplify it further. In the first part, I'll take runtime as an input and give the corresponding title as the output. In the second, I'll find out the maximum runtime of a title and the minimum in the third.

The formula for the first part:

INDEX($A$4:$B$6174,MATCH(J5,B5:B6174,0)+1, 1)

J5 is the cell in which input is entered; range A consists of Titles, and B contains Duration.

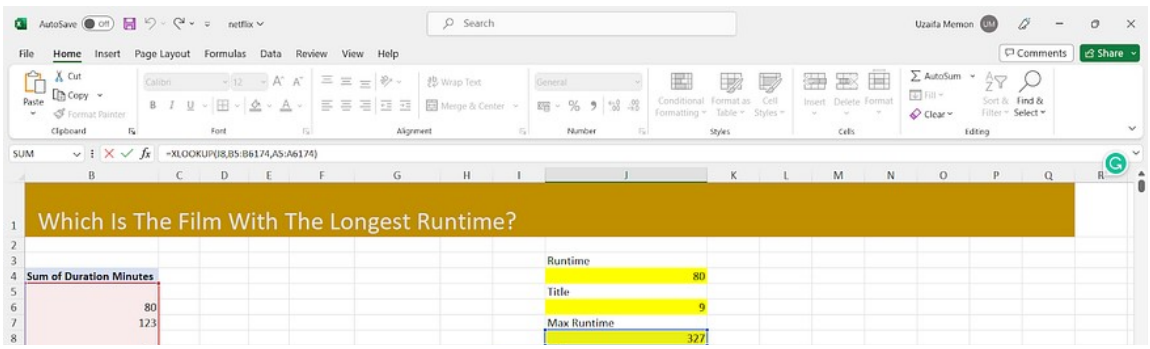
I'm using match to get me the index of the row containing the value mentioned in J5. The index function will then take up the value returned by match and give the corresponding Title as output.

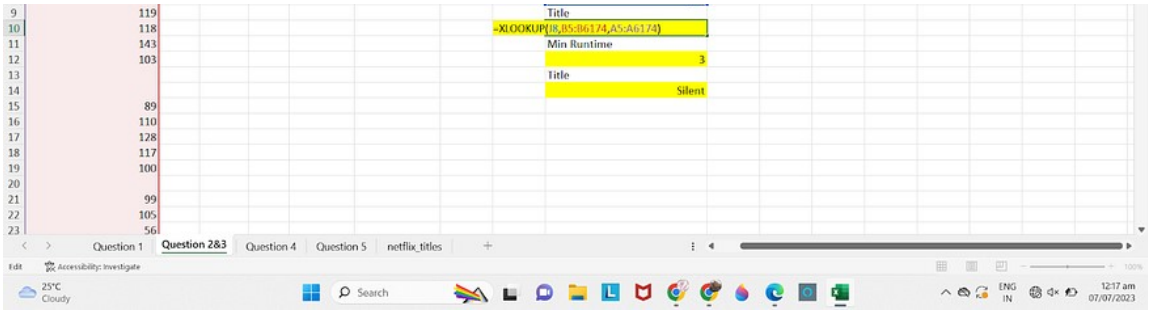
For the second and third parts, functions min and max are used along with xlookup.

Formula:

XLOOKUP(J8,B5:B6174,A5:A6174)

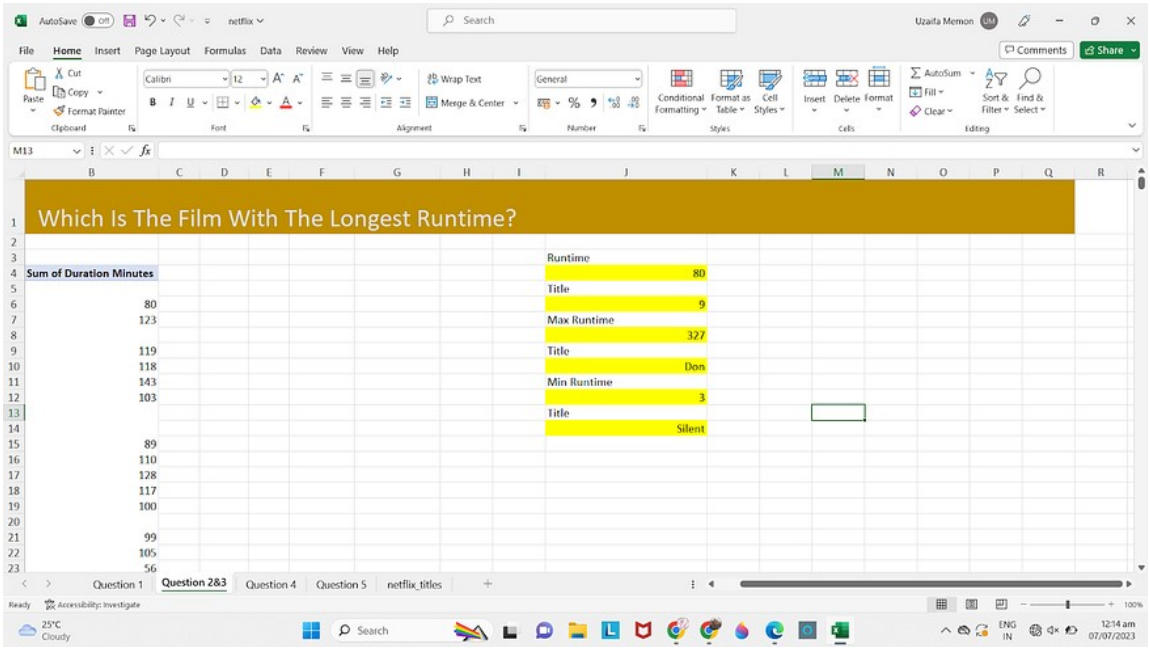
Here, J8 is the cell that contains the maximum runtime value, the second parameter is the array in which the value is to be searched for, and the last is the array from which the corresponding value is to be returned.





Source: Photo by Author

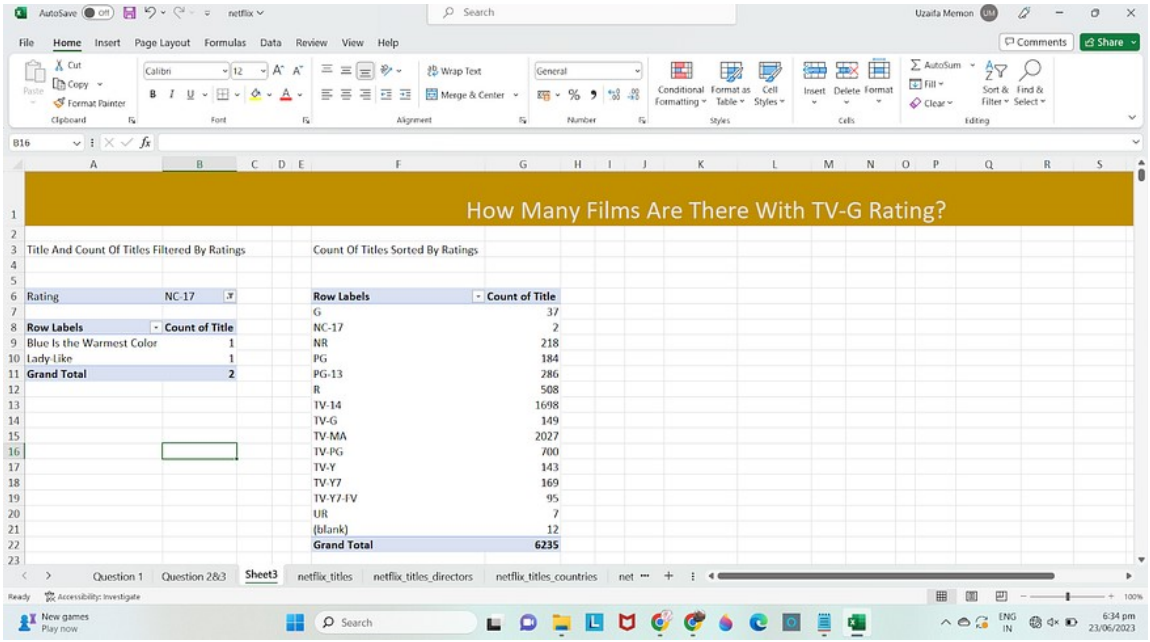
Here's the final output:



Source: Photo by Author

Answer 4

I'm adding a filter (actually, two) to the pivot table for this answer. To do this, open the field list in the PivotTable Analyze ribbon and add the Ratings as a filter. Next, do the same to count the titles sorted by ratings, and you're done!



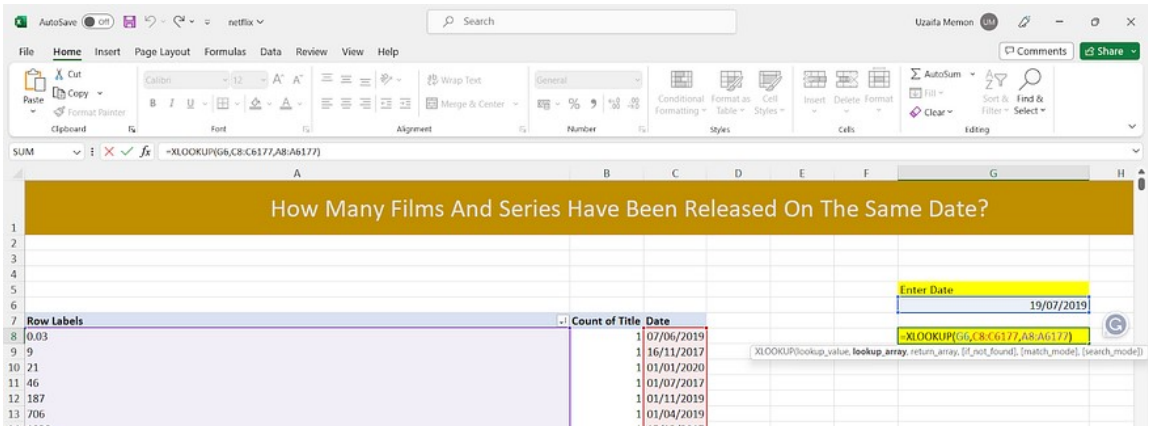
Source: Photo by Author

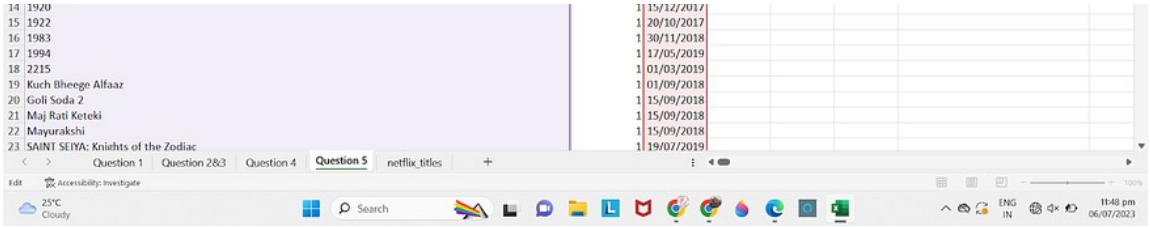
Answer 5

For the last answer, to find the title released given the date, I'll be using Lookup again. This formula is probably my favourite and will soon be yours, too!

The formula used here is where G6 is the cell in which the date is entered as the output.

XLOOKUP(G6,C8:C6177,A8:A6177)





Source: Photo by Author

Conclusion

Okay, so it has been a long discussion (almost 1500 words) and you probably need a long coffee break because I do, too. So, I wouldn't take a lot of your time. Just wanted to let you know that I'm so glad I completed this article, even if it took me quite a long time. I would be even more glad if it helped you.

Stay tuned for more such extensive case studies with interactive dashboards. Have a good day, folks!

P. S: Here's the link to the Excel file you can access.