

INT 217 PROJECT REPORT
(Project Semester August-December 2020)

TITLE OF THE PROJECT

Submitted by

ASHISH

Registration No *11805251*

Programme and Section *BTECH(CSE) and KM073*

Course Code *INT217*

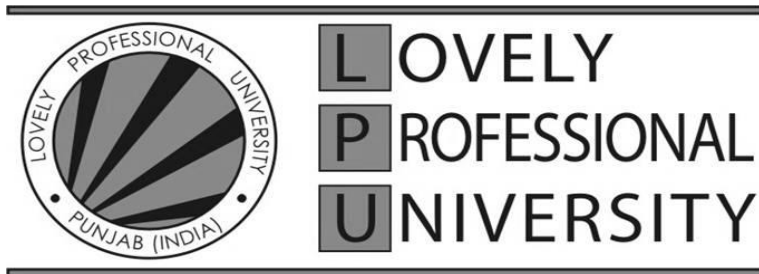
Under the Guidance of

Miss. Ashu U. Id 23631 Assistant Professor

Discipline of CSE/IT

Lovely School of Computer science

Lovely Professional University, Phagwara



DECLARATION

I, **Ashish**, student of **BTECH(CSE)** under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date:

Signature

Registration No. 11805251

Name of the student ASHISH

Acknowledgements

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to *Miss Ashu* for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of Lovel Professional University for their kind co-operation and encouragement which help me in completion of this project.

My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

Table Of Content

1. [Introduction](#)
2. [Objectives/Scope of the Analysis](#)
3. [Source of dataset](#)
4. [ETL process](#)
5. [Analysis on dataset](#)
 - [Total number of movies and tv shows](#)
 - [What is the most rating given?](#)
 - [Which country provides the most content to Netflix?](#)
 - [Top 10 genre in movies and tv shows.](#)
 - [Content added over the month and year](#)
6. [References](#)

Introduction

Netflix, Inc. is an American media-services provider and production company headquartered in Los Gatos, California, founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California.

The company's primary business is its subscription-based streaming service which offers online streaming of a library of films and television programs, including those produced in-house. As of April 2020, Netflix had over 193 million paid subscriptions worldwide, including 73 million in the United States. It is available worldwide except in the following: mainland China (due to local restrictions), Iran, Syria, North Korea, and Crimea (due to U.S. sanctions). The company also has offices in French, United States, United Kingdom, Brazil, the Netherlands, India, Japan, and South Korea. Netflix is a member of the Motion Picture Association of America (MPAA). Today, the company produces and distributes content from countries all over the globe.

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

Objective

There are lot of analysis to do with this dataset. But I will do some 5 analysis, so I am writing their objectives names

1. Total number of movies and tv shows
2. What is the most rating given?
3. Which country provides the most content to Netflix?
4. Top 10 genre in movies and tv shows.
5. Content added over the month and year

Source of Dataset

<https://www.kaggle.com/shivamb/netflix-shows/download>

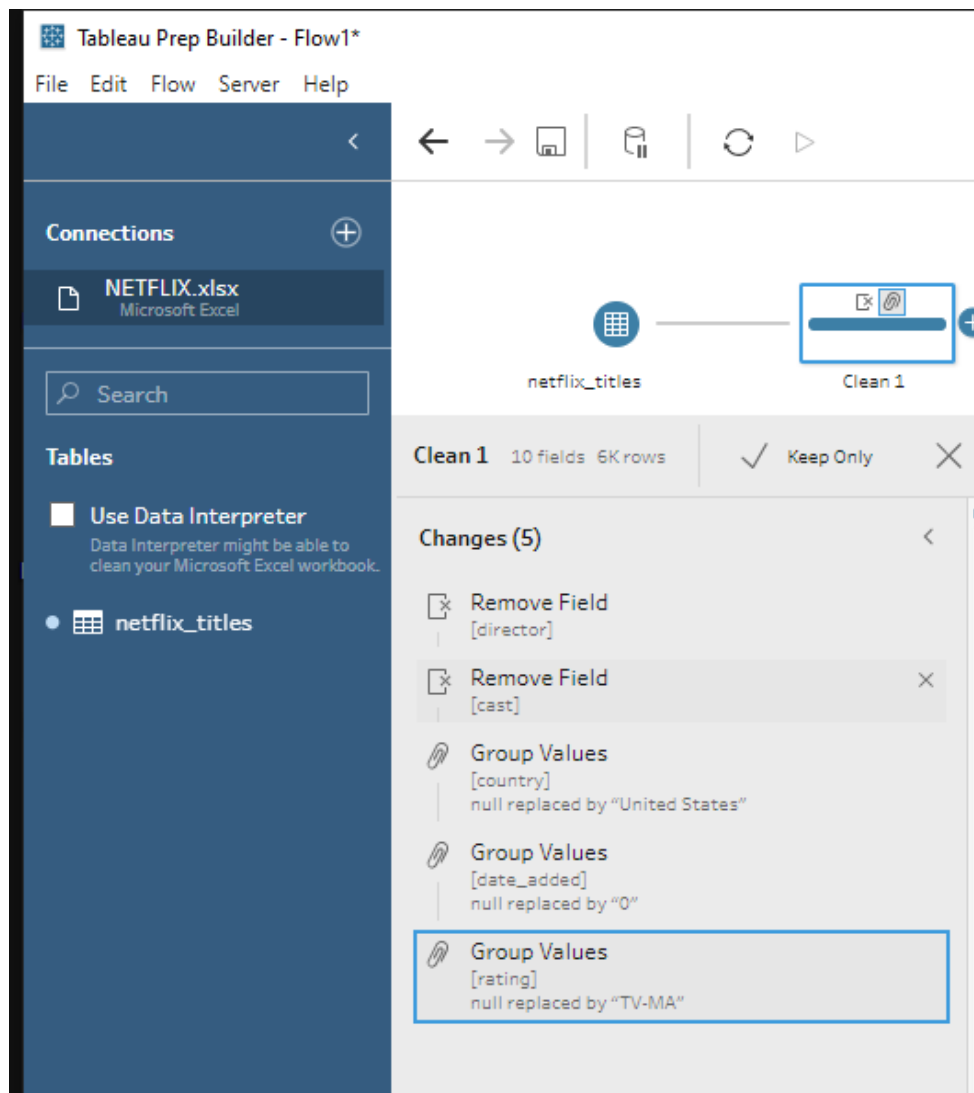
ETL Process

Extraction: - During extraction, the desired data is identified and extracted from many different sources, including database systems and applications.

This dataset has imported from Kaggle.com. Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals.

Transformation: - In this Phase I had done so many transformations.

First, After analysis of this dataset. I have found so many null values, irrelevant data



As you can see in above screen shot. This is tableau Prep Builder a software for which I used it for transforming data. I have dataset named “netflix_titles”. This is the original Dataset from Kaggle. In this dataset I made some changes as you can see above i.e. Changes (5).so let us Know about it one by one

1. I removed “**director**” field because there is so many null values and also by dropping this column doesn’t effect on our visualization as mentioned in the objective.

2. The same changes made to “**Cast**” column because it also has so many null values.
3. After this “**Country**” column also has null values.so instead of dropping the whole column we can just replace the null values in that column with **United States** as there are many shows are aired in the country and Netflix itself is created in that country.
4. Also “date_added” column null values can be replaced by 0.
5. There is very less null value in the rating column so we can replace that with most rated one i.e. **TV-MA**

Now we have handled all the null values.

After this we have date_added column.so we extract only **month** from it because we already have release year column.

So, for getting only month from date_added column I used Excel formula

=TEXT (“Column name”, mmm)

This formula gives first 3 letter of month in text

For e.g. 01-nov-2020 gives “nov”

After this I delete date_added column and save all months in new column named “**month**”

Also deleted show id, description because of no use

Now After all this my final dataset named “FINAL OUTPUT” ready for analysis.

After all changes data set look like below image

AutoSave

Off

Final Output

Search

Ashish

Share

Comments

FileHomeInsertPage LayoutFormulasDataReviewViewHelp

117

This dataset is going to be use for our analysis and visualization.

Analysis on dataset

Let's do our first analysis.

1. Total number of movies and tv shows on Netflix.

Introduction

There are many types of tv shows and movies are present in dataset. Let's find out what type of contents and how many are present in the dataset.

Requirements

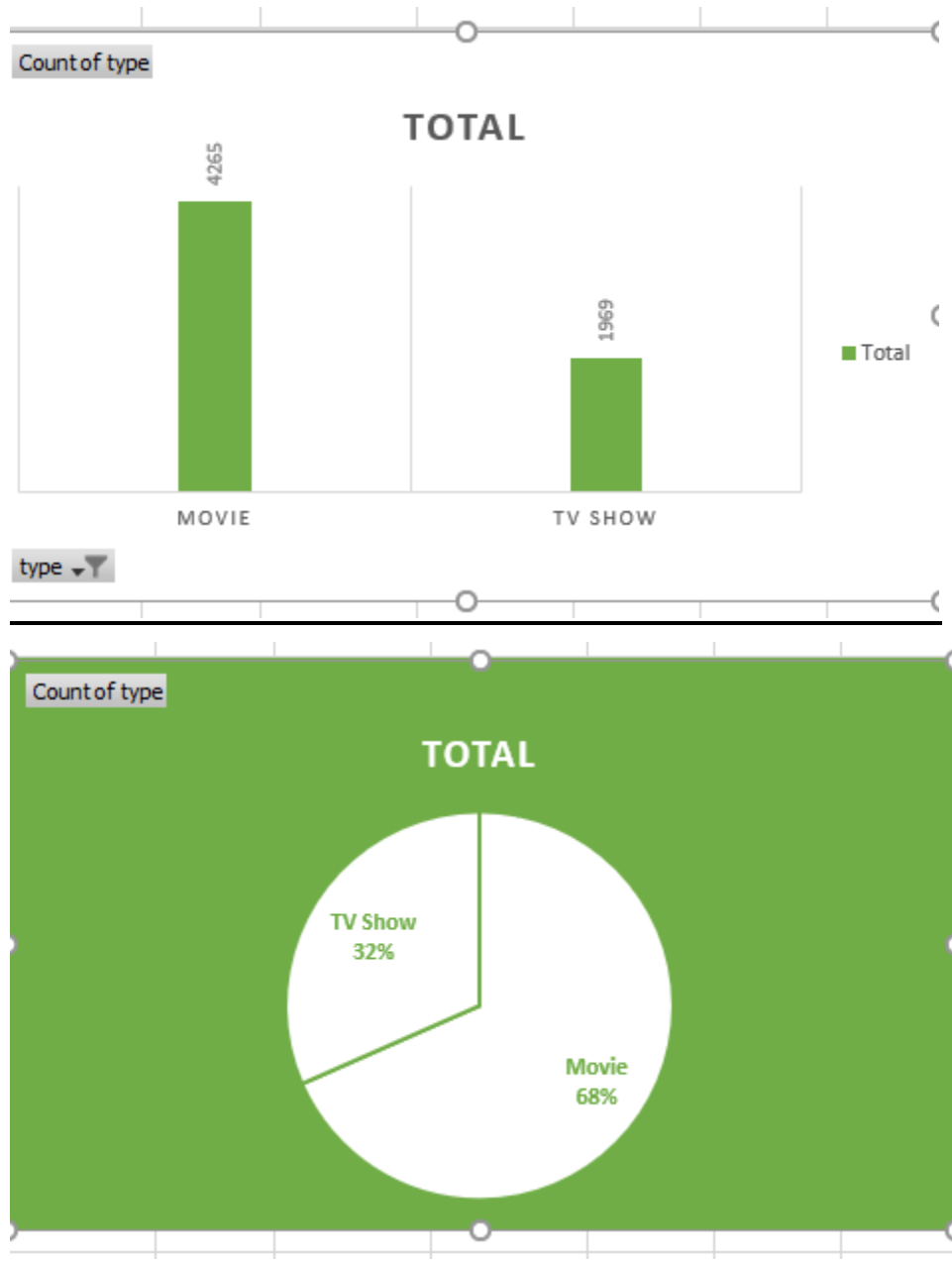
we must convert dataset in to pivot tables. After this select type (have types of contents) in row and drag the field into values for values.

Analysis result

	A	B
1		
2		
3	Row Labels	Count of type
4	Movie	4265
5	TV Show	1969
6	Grand Total	6234
7		
8		

We can see that movies are more in number than Tv shows. Means director are more focusing on movies as because Most people doesn't have time to watch a series continuously so they prefer to watch a movie because it will take just about 90 min.

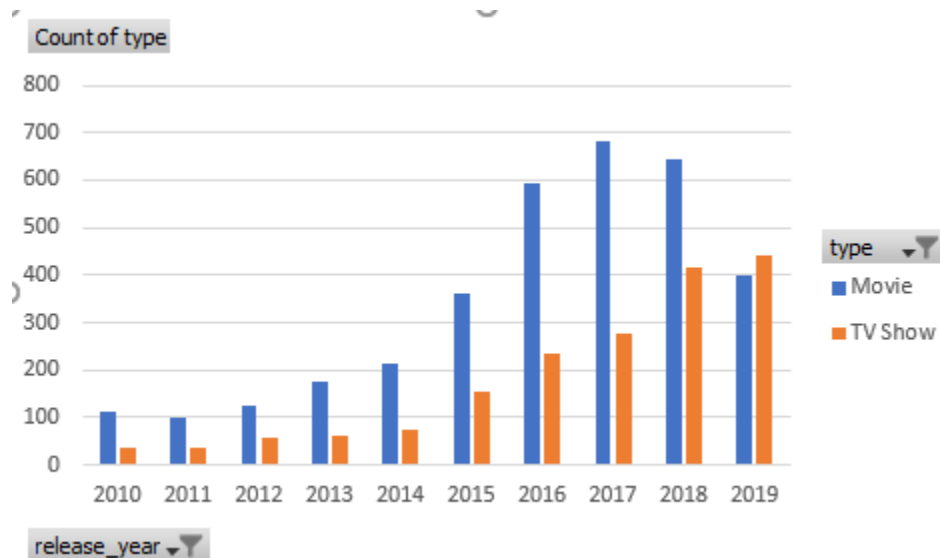
Visualization



Nearly 2/3rd of the content on Netflix are movies and remaining 1/3rd of them are TV Show

If we use type and release year in pivot table for analysis. we get something like this...

Count of type	Column Labels		
Row Labels	Movie	TV Show	Grand Total
2010	111	38	149
2011	100	36	136
2012	125	58	183
2013	177	60	237
2014	213	75	288
2015	363	154	517
2016	593	237	830
2017	682	277	959
2018	646	417	1063
2019	400	443	843
Grand Total	3410	1795	5205



- Netflix is increasing her focusing on Movies rather than TV SHOW in the recent years.
- Netflix try to increase More and More their TV Shows rather Than Movies after 2017.
- This could be because the people who are the most excited about particular movies will have seen them in the cinema before they become available on Netflix, meaning that they have less incentive to check them out on Netflix when the chance comes up.
- TV shows tend to be less expensive to make than movies, meaning that Netflix

2. What is the most rating given on Netflix?

Introduction

The explanation of each content in Netflix:

TV-Y means that this program is generally designed to be viewed by very young audiences under the age of 7 (ages 0 to 6).

TV-Y7 means that a program may not be suitable for children under 7.

TV-G in the United States TV Parental Guidelines signifies content that is suitable for all audiences. Some children's programs that have content that teens or adults will relate to use a TV-G rating, as opposed to a TV-Y rating. This rating is also used for shows with inoffensive content (such as cooking shows, religious programming, nature documentaries, shows about pets and animals, classic television shows, and many shows on Disney Channel carry this rating (particularly sitcoms).

TV-Y7-FV is recommended for ages 7 and older, with the unique advisory that the program contains fantasy violence.

TV-MA are usually created for an adult audience. Some content may not be appropriate for children under the age of 17, due to strong intense violence and particularly coarse language.

TV-14 Parents strongly cautioned. This program contains some material that many parents would find unsuitable for children under 14 years of age.

TV-PG: Parental Guidance Suggested. This program contains material that parents may find unsuitable for younger children.

R Under 17 requires accompanying parent or adult guardian, Parents are urged to learn more about the film before taking their young children with them.

PG-13 Parents Strongly Cautioned, Some Material May Be Inappropriate for Children Under 13.

NR or UR: If a film has not been submitted for a rating or is an uncut version of a film that was submitted.

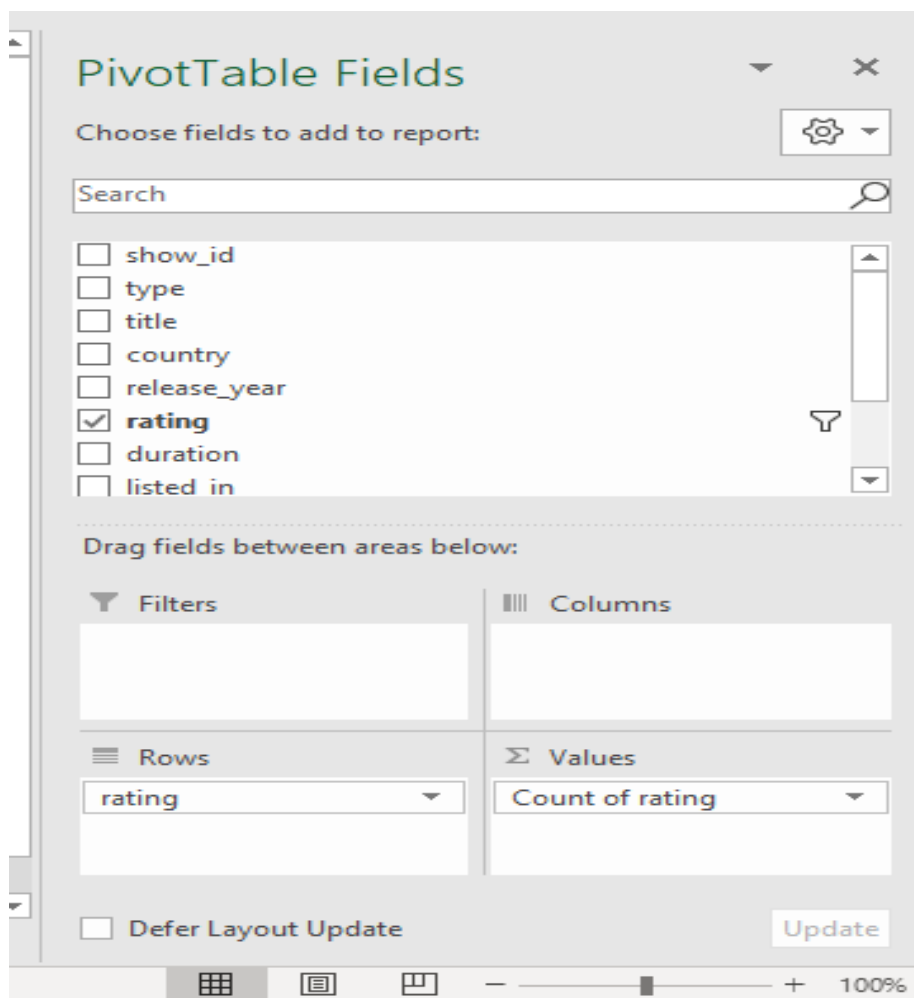
PG: Some material may not be suitable for children, May contain some material parents might not like for their young children.

G: All ages admitted. Nothing that would offend parents for viewing by children.

NC-17 a rating assigned to a movie by the Motion Picture Association of America advising that persons under the age of 18 will not be admitted to a theater showing the film.

Requirements

We required rating column in both “Axes categories” and “values” in pivot table field in excel

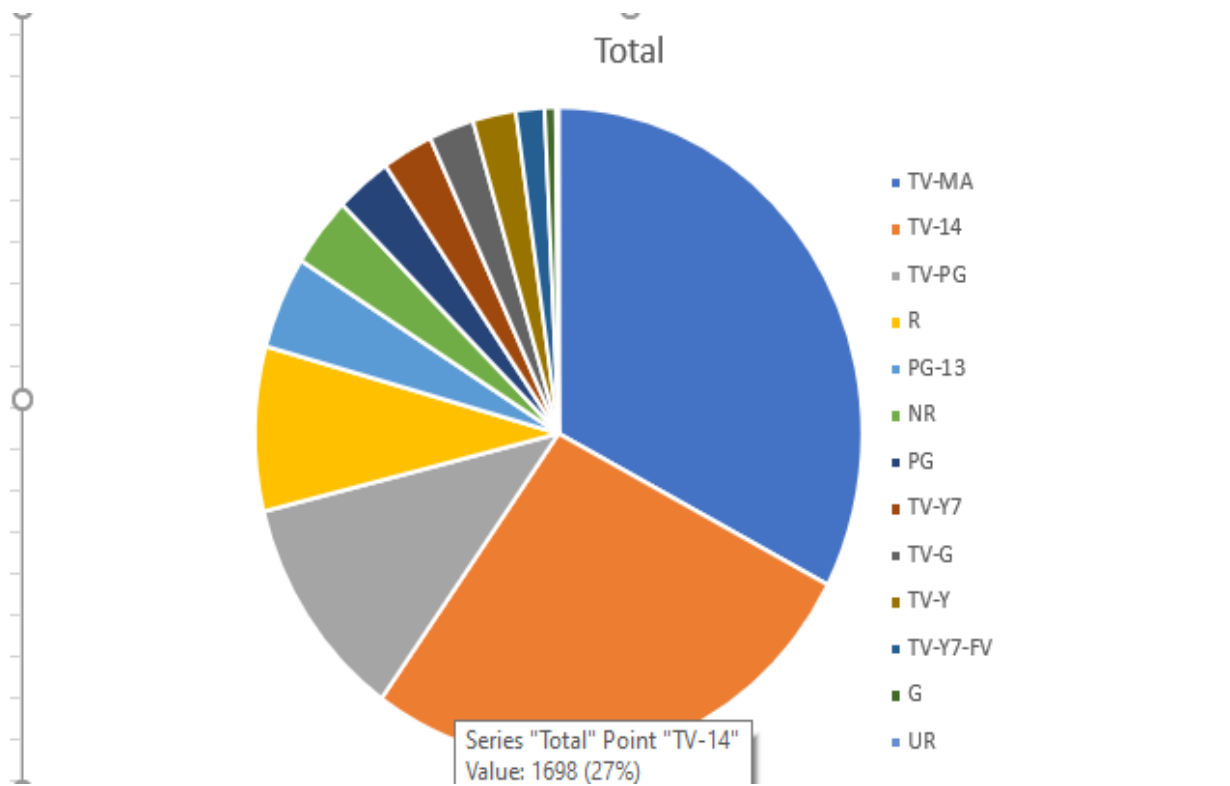
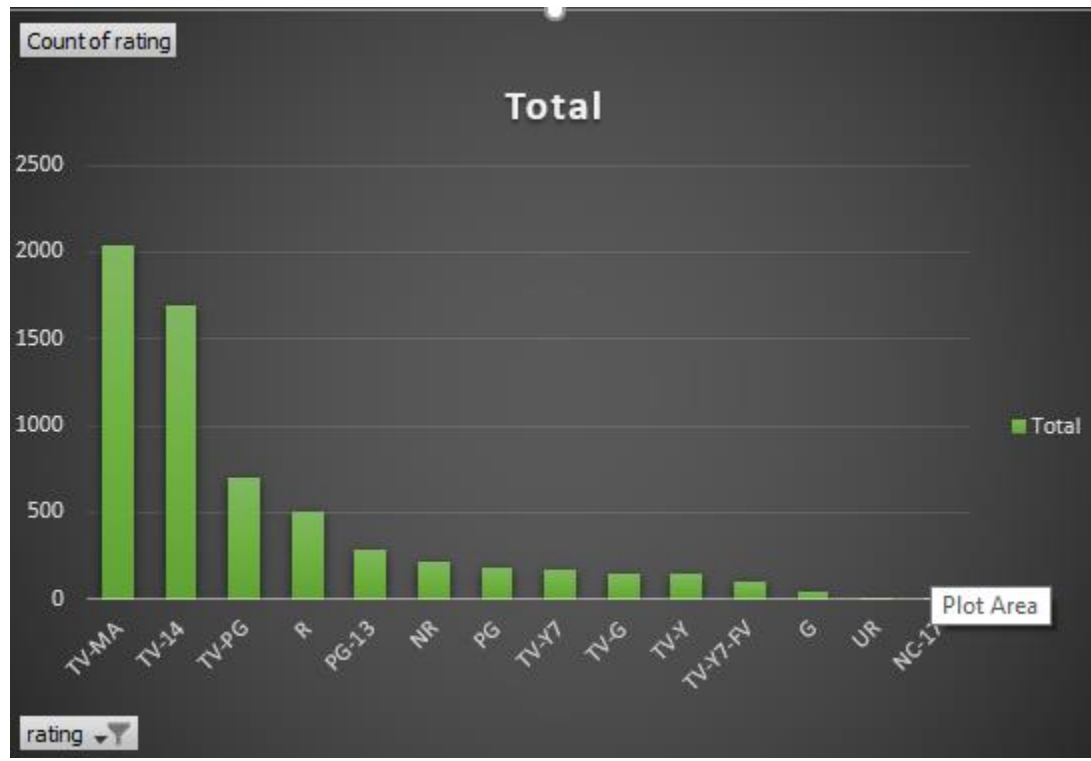


Analysis Result

3	Row Labels	Count of rating
4	G	37
5	NC-17	2
6	NR	218
7	PG	184
8	PG-13	286
9	R	508
10	TV-14	1698
11	TV-G	149
12	TV-MA	2037
13	TV-PG	701
14	TV-Y	143
15	TV-Y7	169
16	TV-Y7-FV	95
17	UR	7
18	Grand Total	6234
19		

Here we can see that the most category audience in Netflix are the adult's audience. "TV-MA" Some content may not be appropriate for children under the age of 17, due to strong intense violence and particularly coarse language.

Visualization



3. Which country provides the most content to Netflix

Introduction


Netflix is available in 190 different countries, but the catalog of film and TV shows is different depending on location. Ever wonder which country has the biggest and best collection of content?


Since there are contents that are produced in different countries, so let's check that by our analysis


Requirements

We need country column in row and count of country in values


PivotTable Fields


Choose fields to add to report: 

Search 

☐ type 

☐ title


☒ **country** 

☐ release_year 



☐ rating





☐ duration


☐ listed_in





☐ description 

Drag fields between areas below:

 Filters	 Columns
<div></div>	<div></div>

 Rows	 Values
<div>country </div>	<div>Count of country </div>
<div></div>	<div></div>

☐ Defer Layout Update 

   -  + 100%

Analysis Result

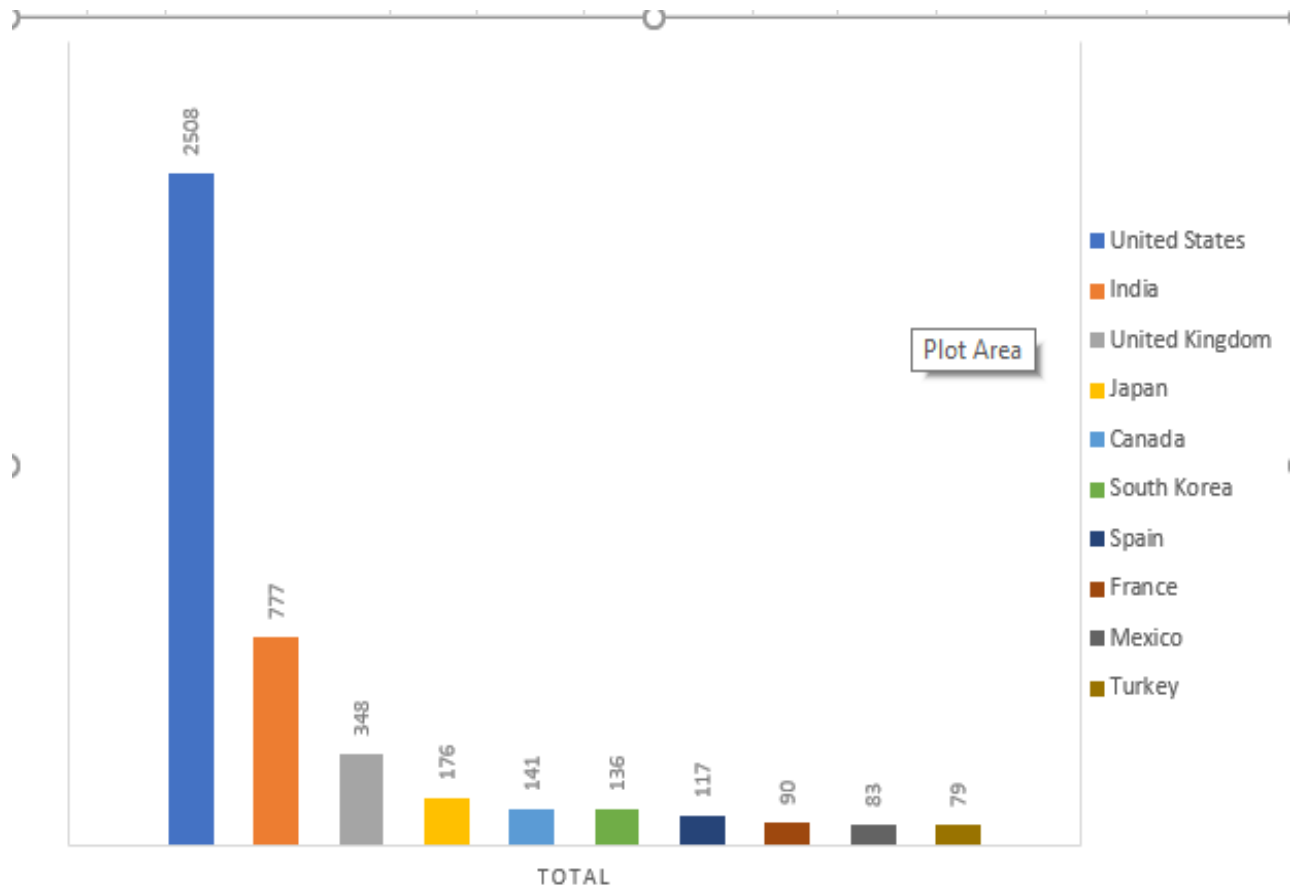
Here are top 10 countries have content on Netflix

Row Labels	Count of country
Canada	141
France	90
India	777
Japan	176
Mexico	83
South Korea	136
Spain	117
Turkey	79
United Kingdom	348
United States	2508
Grand Total	4455

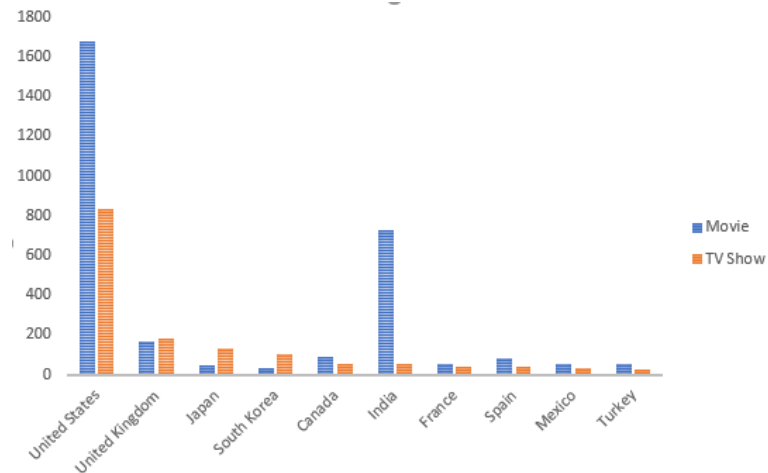
Here is another top 10 analysis with TV shows and movies list according to country

Row Labels	Movie	TV Show
Canada	88	53
France	50	40
India	724	53
Japan	47	129
Mexico	54	29
South Korea	32	104
Spain	80	37
Turkey	55	24
United Kingdom	170	178
United States	1677	831

Visualization



We can see that after united states, India is second largest contributor to Netflix



- India is more focusing on releasing movies on Netflix

4. Top 10 genre in movies and tv shows.

Introduction

A genre film is one that is easily categorized film accepted film and TV genres based on similarities in narrative elements or the emotional response to the piece of work. That means these works are judged by who's in them, how they are shot, and aspects of their screenplays.

Top genre in movies show are:

Drama

Comedy

Sci-fi

Action

Adventure

Animation

Mystery

Crime

Thriller

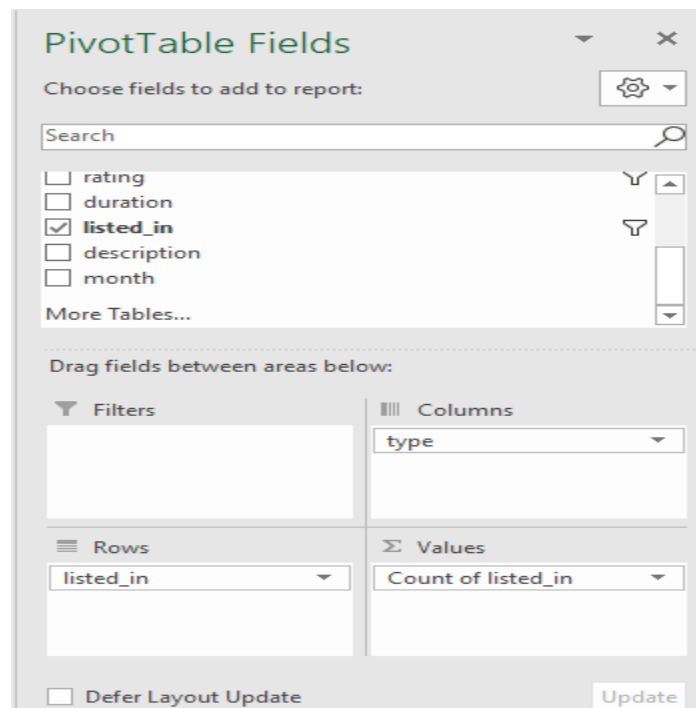
Romance

Horror

Etc.

Requirements

We required “listed_in” column and “type” Column for this analysis.

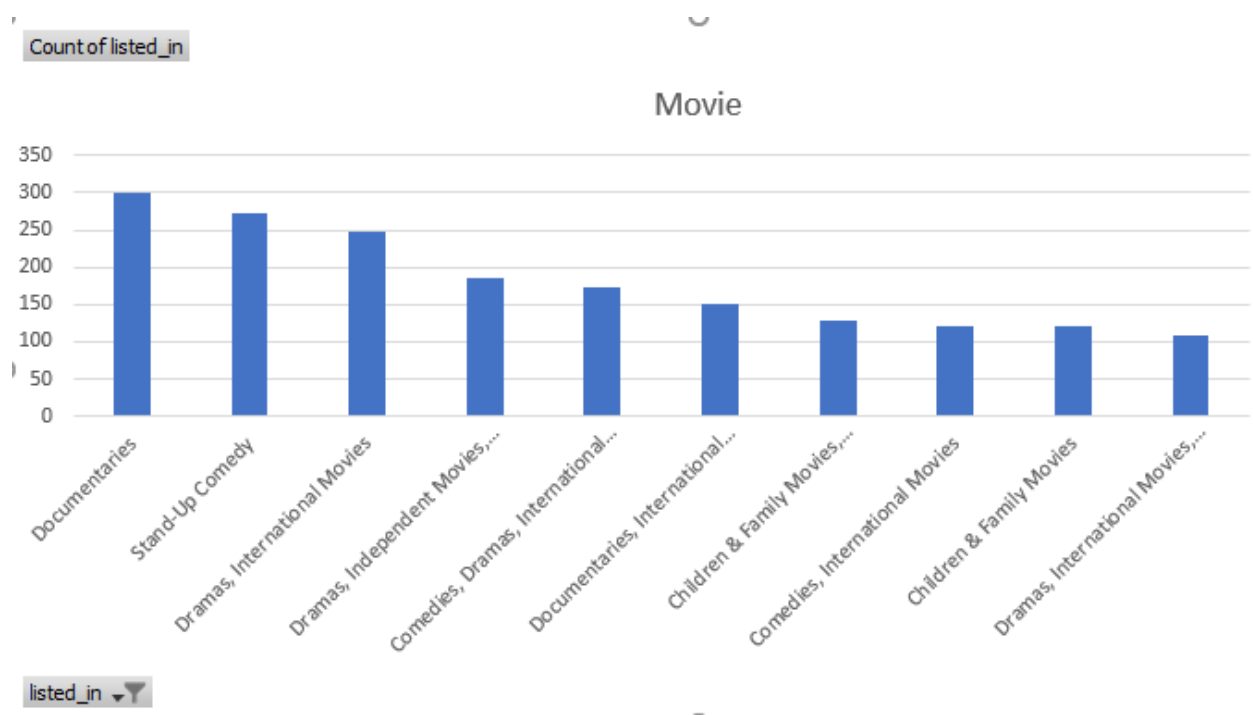


Analysis Result

Top 10 genre in movies.....

Count of listed_in	Column Labels
Row Labels	Movie
Documentaries	299
Stand-Up Comedy	273
Dramas, International Movies	248
Dramas, Independent Movies, International Movies	186
Comedies, Dramas, International Movies	174
Documentaries, International Movies	150
Children & Family Movies, Comedies	129
Comedies, International Movies	120
Children & Family Movies	120
Dramas, International Movies, Romantic Movies	108
Grand Total	1807

Visualization

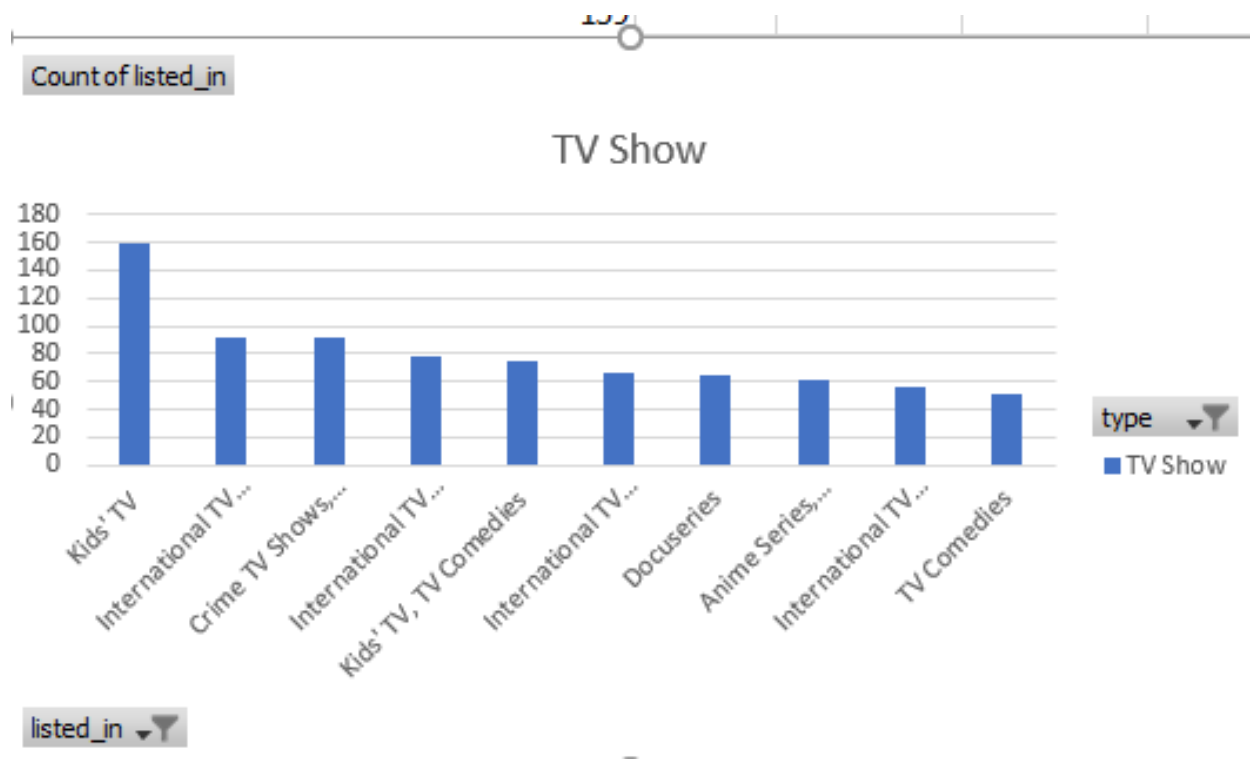


Analysis result

Top 10 genre in Tv Shows

3	Count of listed_in	Column Labels
4	Row Labels	TV Show
5	Kids' TV	159
6	International TV Shows, TV Dramas	92
7	Crime TV Shows, International TV Shows, TV Dramas	92
8	International TV Shows, Romantic TV Shows, TV Dramas	78
9	Kids' TV, TV Comedies	75
10	International TV Shows, Romantic TV Shows, TV Comedies	66
11	Docuseries	65
12	Anime Series, International TV Shows	62
13	International TV Shows, Korean TV Shows, Romantic TV Shows	56
14	TV Comedies	52
15	Grand Total	797
16		

Visualization



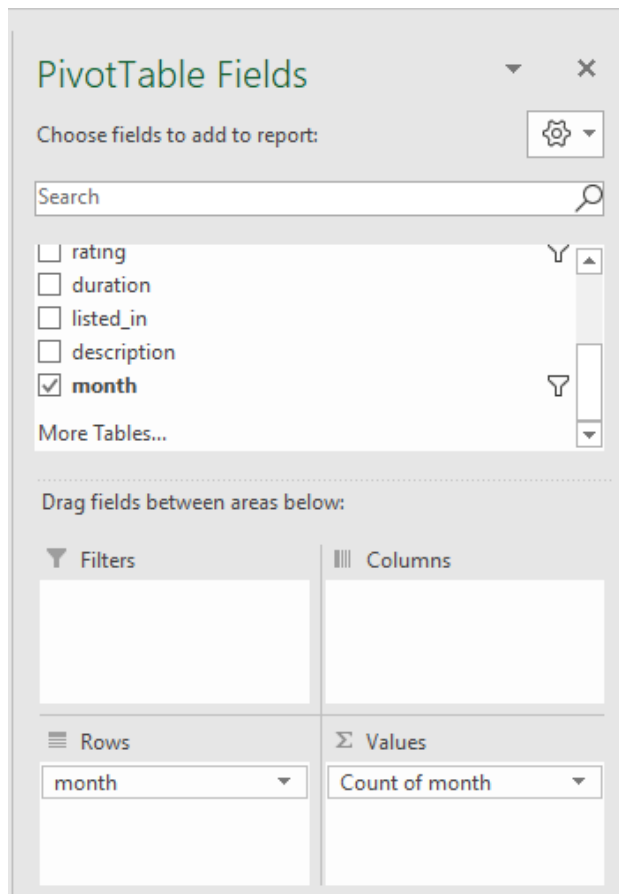
5. Content added over the month and year

Introduction

Netflix release new movies and tv shows each month. So here we are going to analysis this.

Requirements

We need month and year and count of year and month for this analysis. Like this...

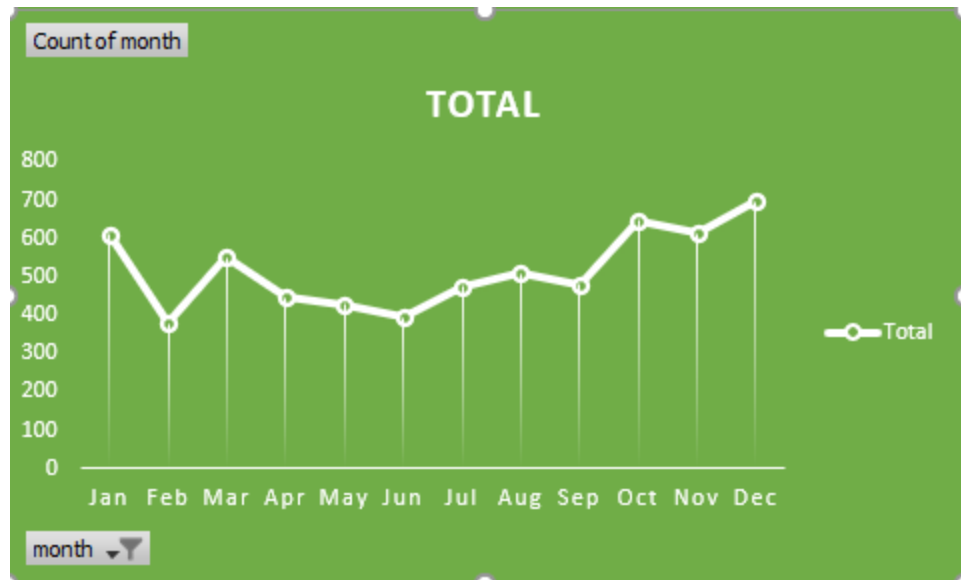


Analysis result

Content added over the month

3	Row Labels	Count of month
4	Jan	610
5	Feb	378
6	Mar	551
7	Apr	447
8	May	428
9	Jun	393
10	Jul	474
11	Aug	509
12	Sep	479
13	Oct	646
14	Nov	612
15	Dec	696
16	(blank)	
17	Grand Total	6223

Visualization



The growth in contents are higher in the first three months and the last three months of the year.

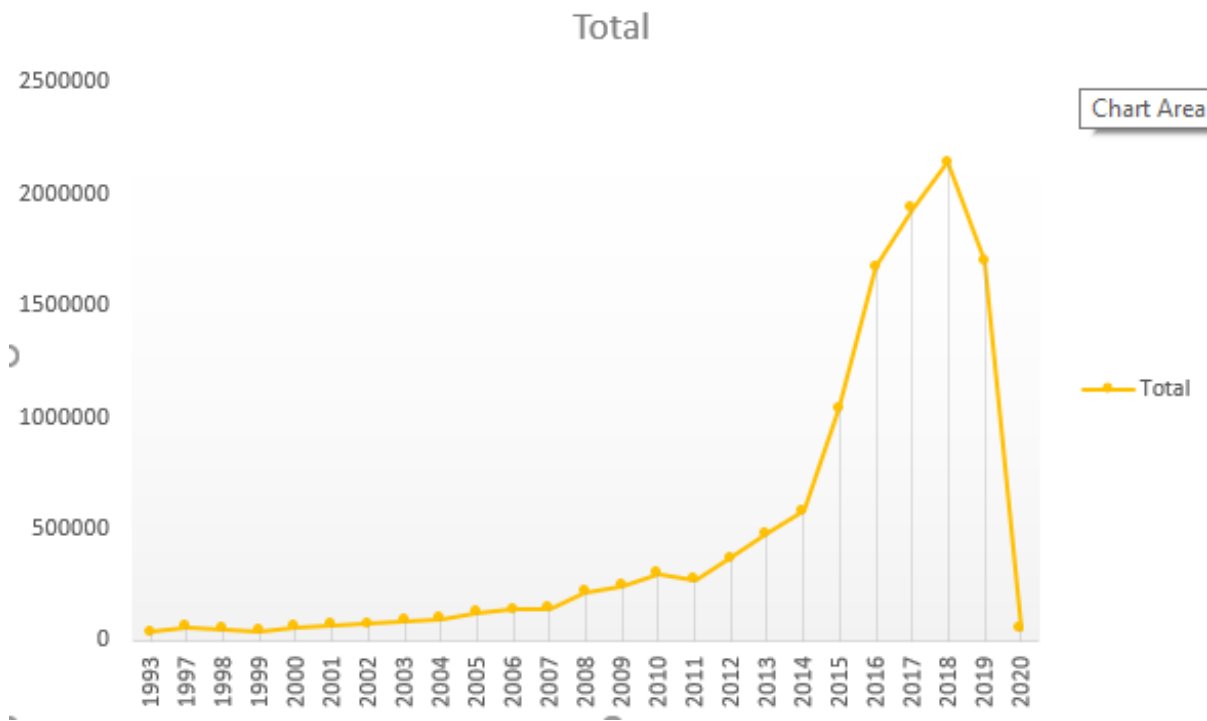
Least number of contents are added in the month of February

Analysis Result.

Content added over the year

3	Row Labels	Sum of release_year
4	1993	37867
5	1997	61907
6	1998	51948
7	1999	41979
8	2000	62000
9	2001	68034
10	2002	76076
11	2003	86129
12	2004	98196
13	2005	126315
14	2006	136408
15	2007	142497
16	2008	214856
17	2009	243089
18	2010	299490
19	2011	273496
20	2012	368196
21	2013	477081
22	2014	580032
23	2015	1041755
24	2016	1673280
25	2017	1934303
26	2018	2145134
27	2019	1702017
28	2020	50500
29	Grand Total	11992585

- Visualization The growth in content started from 2013



References

- [Kaggle](#)
- [Wikipedia](#)
- [Google](#)