Introduction to Data Management Project Report

(August – December 2021)

PROJECT REPORT

ON

Analysing Movies On Netflix, Prime, Hulu, Disney+

Submitted by

YS Pavan Kumar

11901777

Programme : Bachelor of Technology

Section : KM008

Course Code : INT217

Discipline of CSE/IT

Lovely School of Computer Science and engineering

Lovely Professional University , Phagwara

DECLARATION

I, **YS Pavan Kumar** , student of Computer Science & Engineering 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

Name of the student : **YS Pavan Kumar**

Registration number **: 11901777**

ACKNOWLEDGEMENT

Firstly , I Would like to take this opportunity to express my gratitude towards my mentor for the subject INT217 Ms. Komal Arora and Lovely Professional University for the constant support and guidance throughout.
I Would also like to thank my friends and class mates for the constant support which motivated me to complete the project in due time

YS Pavan Kumar

CONTENTS

INTRODUCTION

• Data Management is a set of disciplines and techniques used to process, store, organize and analyse the data

•Data analysis is a process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, informing conclusions and supporting decision making

•Data management can increase the visibility of your organization's data Assets making it easier for people to quickly and confidently find the rightdata for their analysis. Data visibility allows your company to be more organized and productive, allowing employees to find the data they need to better do their jobs.

•This project is based on such data analysis on Movies in Netflix, prime, Hulu, Disney+

•Data Analysis Consists of the following phases :

Data Requirement Gathering

Data Collection

Data Cleaning

Data Analysis

Data Interpretation

Data visualization

•Netflix, Prime, Hulu, Disney+ are online streaming platforms for movies , web series and different programmes

•This data set contains 12 fields

OBJECTIVES

After analysis of the data set the aim of this project is to give answer of given objectives

Objectives

- Number of movies released in each year
- Top 10 rated movies according to IMDB
- Number of Movies released per country
- Targeted age group of movies
- Top 10 Rated Movies according to Rotten tomatoes
- Longest Run Time Movies
- Number of movies released in Netflix, Hulu, Prime, Disney+

Source

Source of Data Set : https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

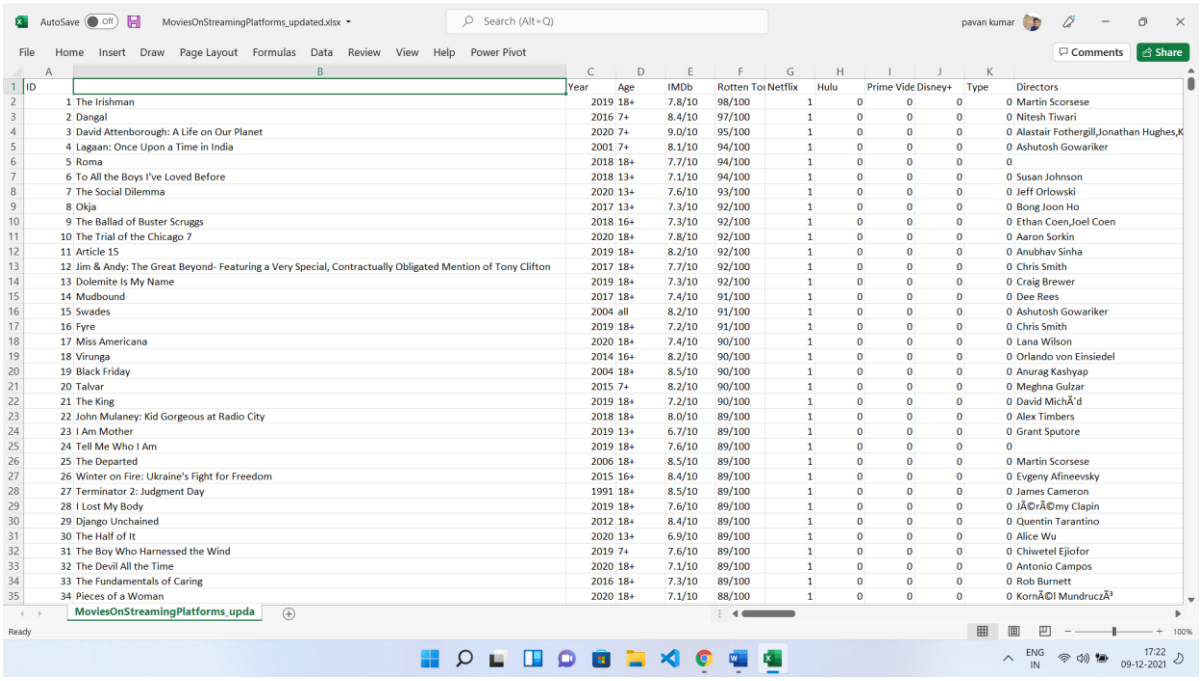
Kaggle :

Kaggle is an Airbnb for Data Scientists – this is where they spend their nights and weekends. It's a crowd-sourced platform to attract, nurture, train, and challenge data scientists from all around the world to solve data science, machine learning and predictive analytics problems. It has over 536,000 active members from 194 countries, and it receives close to 150,000 submissions per month. Started from Melbourne, Australia Kaggle moved to Silicon Valley in 2011, raised some 11 million dollars, then ultimately been acquired by the Google in March of2017. Kaggle is the number one stop for data science enthusiasts all around the world who compete for prizes and boost their Kaggle rankings. There are only 94 Kaggle Grandmasters in the world to this date. Kaggle enables data scientists and other developers to engage in running machine learning contests, write and share code, and to host datasets. The types of data science problems posted on Kaggle can be anything from attempting to predict cancer occurrence by examining patient records to analyzing sentiment to evoke by movie reviews and how this affects audience reaction. Different sources post projects on this trailblazing platform. While some are just for educational purposes and fun brain exercises, others are genuine issues that companies are trying to solve. Kaggle the environment competitive by awarding prizes and rankings for winners and participants. The prizes are not only monetary but can also include attractive rewards such as jobs or free products

from the company hosting the competition. Interesting and challenging projects where contributors can Competitions involve solving challenging and interesting problems. Companies post projects to numerous contributors. It especially a great place for beginners who are just trying to break into the data science field. Aside from the competitions that are open to the general public, Kaggle also has private competitions which are only open to top rated participants (Kaggle Masters).Insightful discussions with Industry leaders and learned experts apart from the projects, Kaggle also consists of live discussions between numerous people on the platform. Such forums are very interesting, stimulating, and informative. Through these discussions, you can either seek advice from others or offer advice to people who are dealing with issues you understand Kaggle offers its audience a chance to get into the biggest data science community in the world. This platform is trusted by some of the largest data science companies of the world such as Walmart, Facebook, and Winton Capital. On Kaggle, data scientists get exposure and a chance to work on problems faced by big companies in real-time. While it is not a guarantee, there is always the chance that the company will be impressed enough to recruit.

The host of the competition is in-charge of preparing the data and preparing a detailed description of the problem at hand. To make it more convenient for hosts, Kaggle offers an additional consulting service that can help prepare data and describe the problem in the best possible format.

Data Set :



ID		Year	Age	IMDb	Rotten Tor	Netflix	Hulu	Prime Vide	Disney+	Type	Directors
1	The Irishman	2019	18+	7.8/10	98/100	1	0	0	0	0	Martin Scorsese
2	Dangal	2016	7+	8.4/10	97/100	1	0	0	0	0	Nitesh Tiwari
3	David Attenborough: A Life on Our Planet	2020	7+	9.0/10	95/100	1	0	0	0	0	Alastair Fothergill,Jonathan Hughes,K
4	Lagaan: Once Upon a Time in India	2001	7+	8.1/10	94/100	1	0	0	0	0	Ashutosh Gowariker
5	Roma	2018	18+	7.7/10	94/100	1	0	0	0	0	
6	To All the Boys I've Loved Before	2018	13+	7.1/10	94/100	1	0	0	0	0	Susan Johnson
7	The Social Dilemma	2020	13+	7.6/10	93/100	1	0	0	0	0	Jeff Orlowski
8	Okja	2017	13+	7.3/10	92/100	1	0	0	0	0	Bong Joon Ho
9	The Ballad of Buster Scruggs	2018	16+	7.3/10	92/100	1	0	0	0	0	Ethan Coen,Joel Coen
10	The Trial of the Chicago 7	2020	18+	7.8/10	92/100	1	0	0	0	0	Aaron Sorkin
11	Article 15	2019	18+	8.2/10	92/100	1	0	0	0	0	Anubhav Sinha
12	Jim & Andy: The Great Beyond- Featuring a Very Special, Contractually Obligated Mention of Tony Clifton	2017	18+	7.7/10	92/100	1	0	0	0	0	Chris Smith
13	Dolemite Is My Name	2019	18+	7.3/10	92/100	1	0	0	0	0	Craig Brewer
14	Mudbound	2017	18+	7.4/10	91/100	1	0	0	0	0	Dee Rees
15	Swades	2004	all	8.2/10	91/100	1	0	0	0	0	Ashutosh Gowariker
16	Fyre	2019	18+	7.2/10	91/100	1	0	0	0	0	Chris Smith
17	Miss Americana	2020	18+	7.4/10	90/100	1	0	0	0	0	Lana Wilson
18	Virunga	2014	16+	8.2/10	90/100	1	0	0	0	0	Orlando von Einsiedel
19	Black Friday	2004	18+	8.5/10	90/100	1	0	0	0	0	Anurag Kashyap
20	Talvar	2015	7+	8.2/10	90/100	1	0	0	0	0	Meghna Gulzar
21	The King	2019	18+	7.2/10	90/100	1	0	0	0	0	David MichÃ´d
22	John Mulaney: Kid Gorgeous at Radio City	2018	18+	8.0/10	89/100	1	0	0	0	0	Alex Timbers
23	I Am Mother	2019	13+	6.7/10	89/100	1	0	0	0	0	Grant Sputore
24	Tell Me Who I Am	2019	18+	7.6/10	89/100	1	0	0	0	0	
25	The Departed	2006	18+	8.5/10	89/100	1	0	0	0	0	Martin Scorsese
26	Winter on Fire: Ukraine's Fight for Freedom	2015	16+	8.4/10	89/100	1	0	0	0	0	Evgeny Afineevsky
27	Terminator 2: Judgment Day	1991	18+	8.5/10	89/100	1	0	0	0	0	James Cameron
28	I Lost My Body	2019	18+	7.6/10	89/100	1	0	0	0	0	JÃ©rÃ©my Clapin
29	Django Unchained	2012	18+	8.4/10	89/100	1	0	0	0	0	Quentin Tarantino
30	The Half of It	2020	13+	6.9/10	89/100	1	0	0	0	0	Alice Wu
31	The Boy Who Harnessed the Wind	2019	7+	7.6/10	89/100	1	0	0	0	0	Chiwetel Ejiofor
32	The Devil All the Time	2020	18+	7.1/10	89/100	1	0	0	0	0	Antonio Campos
33	The Fundamentals of Caring	2016	18+	7.3/10	89/100	1	0	0	0	0	Rob Burnett
34	Pieces of a Woman	2020	18+	7.1/10	88/100	1	0	0	0	0	KornÃ©l MundruczÃ³

ETL PROCESS:

ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system.

Full form of ETL is Extract, Transform and Load. The triple combination of ETL provides crucial functions that are many times combined into a single application or suite of tools that help in the following areas:

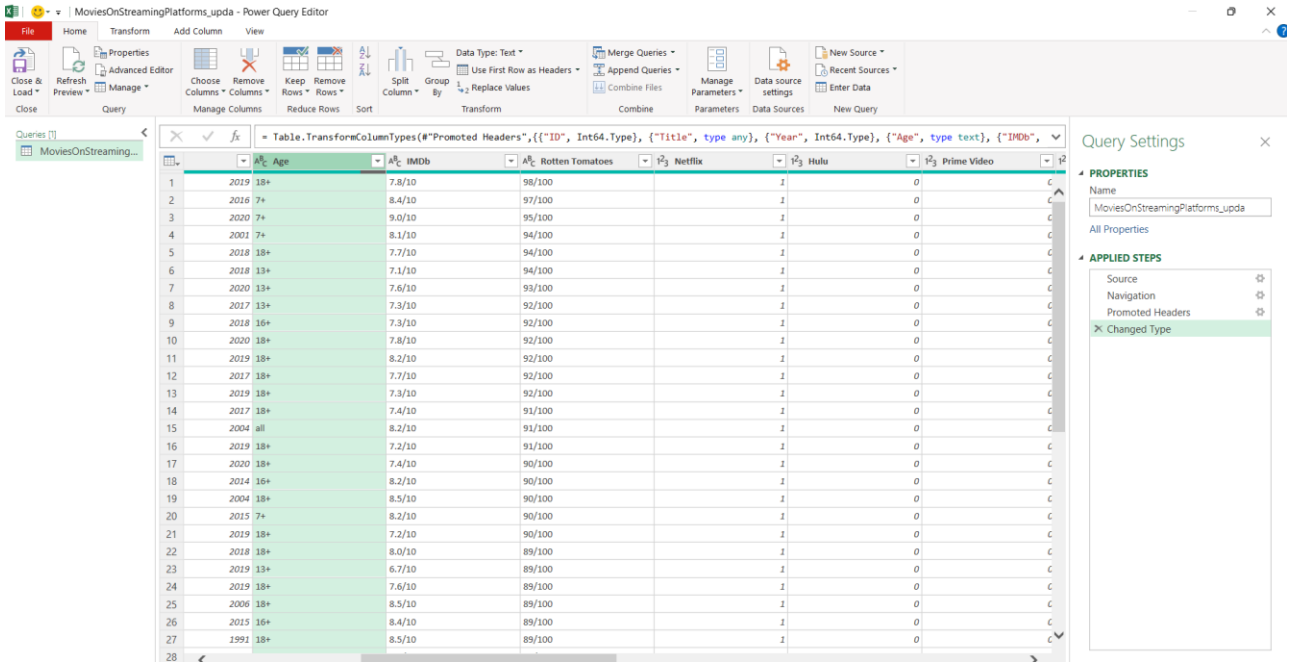
- Enhances Business Intelligence solutions for decision making.
- Allows verification of data transformation, aggregation, and calculations rules.
- Allows sample data comparison between source and target system.
- Helps to improve productivity as it codifies and reuses without additional technical skills.

Initially, the raw dataset was arranged as shown in given picture

	ID		Year	Age	IMDb	Rotten To	Netflix	Hulu	Prime Vide	Disney+	Type	Directors
2	1	The Irishman	2019	18+	7.8/10	98/100	1	0	0	0	0	Martin Scorsese
3	2	Dangal	2016	7+	8.4/10	97/100	1	0	0	0	0	Nitesh Tiwari
4	3	David Attenborough: A Life on Our Planet	2020	7+	9.0/10	95/100	1	0	0	0	0	Alastair Fothergill,Jonathan Hughes,K
5	4	Lagaan: Once Upon a Time in India	2001	7+	8.1/10	94/100	1	0	0	0	0	Ashutosh Gowariker
6	5	Roma	2018	18+	7.7/10	94/100	1	0	0	0	0	
7	6	To All the Boys I've Loved Before	2018	13+	7.1/10	94/100	1	0	0	0	0	Susan Johnson
8	7	The Social Dilemma	2020	13+	7.6/10	93/100	1	0	0	0	0	Jeff Orlowski
9	8	Okja	2017	13+	7.3/10	92/100	1	0	0	0	0	Bong Joon Ho
10	9	The Ballad of Buster Scruggs	2018	16+	7.3/10	92/100	1	0	0	0	0	Ethan Coen,Joel Coen
11	10	The Trial of the Chicago 7	2020	18+	7.8/10	92/100	1	0	0	0	0	Aaron Sorkin
12	11	Article 15	2019	18+	8.2/10	92/100	1	0	0	0	0	Anubhav Sinha
13	12	Jim & Andy: The Great Beyond- Featuring a Very Special, Contractually Obligated Mention of Tony Clifton	2017	18+	7.7/10	92/100	1	0	0	0	0	Chris Smith
14	13	Dolemite Is My Name	2019	18+	7.3/10	92/100	1	0	0	0	0	Craig Brewer
15	14	Mudbound	2017	18+	7.4/10	91/100	1	0	0	0	0	Dee Rees
16	15	Swades	2004	all	8.2/10	91/100	1	0	0	0	0	Ashutosh Gowariker
17	16	Fyre	2019	18+	7.2/10	91/100	1	0	0	0	0	Chris Smith
18	17	Miss Americana	2020	18+	7.4/10	90/100	1	0	0	0	0	Lana Wilson
19	18	Virunga	2014	16+	8.2/10	90/100	1	0	0	0	0	Orlando von Einsiedel
20	19	Black Friday	2004	18+	8.5/10	90/100	1	0	0	0	0	Anurag Kashyap
21	20	Talvar	2015	7+	8.2/10	90/100	1	0	0	0	0	Meghna Gulzar
22	21	The King	2019	18+	7.2/10	90/100	1	0	0	0	0	David MichÃ´d
23	22	John Mulaney: Kid Gorgeous at Radio City	2018	18+	8.0/10	89/100	1	0	0	0	0	Alex Timbers
24	23	I Am Mother	2019	13+	6.7/10	89/100	1	0	0	0	0	Grant Sputore
25	24	Tell Me Who I Am	2019	18+	7.6/10	89/100	1	0	0	0	0	
26	25	The Departed	2006	18+	8.5/10	89/100	1	0	0	0	0	Martin Scorsese
27	26	Winter on Fire: Ukraine's Fight for Freedom	2015	16+	8.4/10	89/100	1	0	0	0	0	Evgeny Afineevsky
28	27	Terminator 2: Judgment Day	1991	18+	8.5/10	89/100	1	0	0	0	0	James Cameron
29	28	I Lost My Body	2019	18+	7.6/10	89/100	1	0	0	0	0	JÃ©rÃ©my Clapin
30	29	Django Unchained	2012	18+	8.4/10	89/100	1	0	0	0	0	Quentin Tarantino
31	30	The Half of It	2020	13+	6.9/10	89/100	1	0	0	0	0	Alice Wu
32	31	The Boy Who Harnessed the Wind	2019	7+	7.6/10	89/100	1	0	0	0	0	Chiwetel Ejiofor
33	32	The Devil All the Time	2020	18+	7.1/10	89/100	1	0	0	0	0	Antonio Campos
34	33	The Fundamentals of Caring	2016	18+	7.3/10	89/100	1	0	0	0	0	Rob Burnett
35	34	Pieces of a Woman	2020	18+	7.1/10	88/100	1	0	0	0	0	KornÃ©l MundruczÃ³

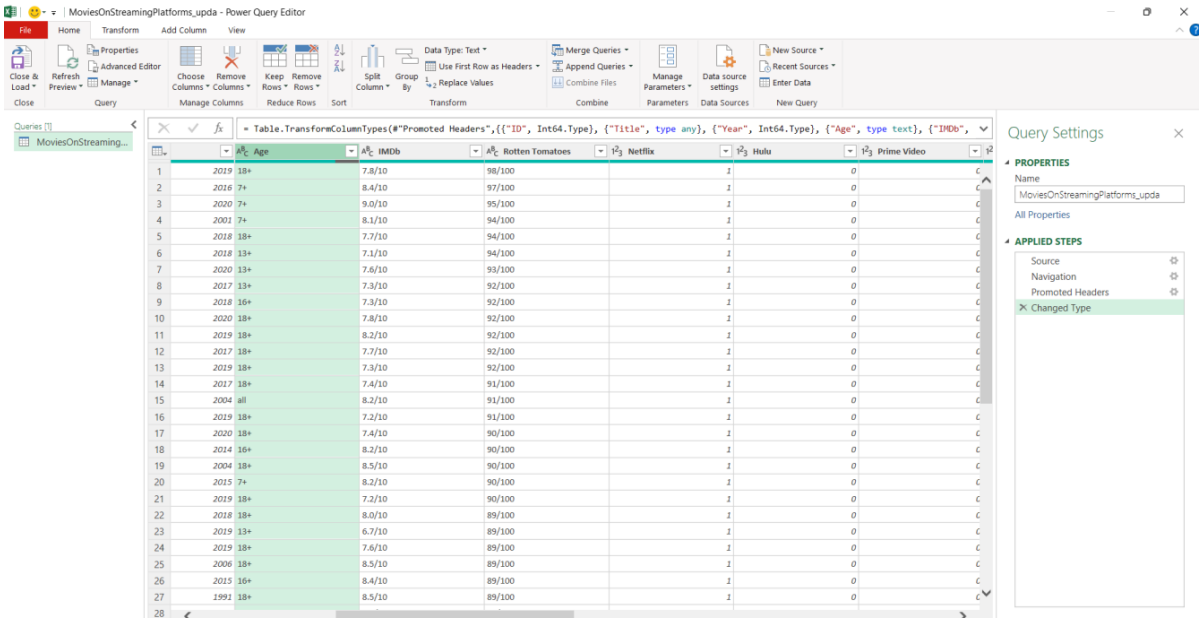
Steps taken to clean dataset thorough ETL process

1) Open blank excel file go to data tab and select get data and select from text/csv and extract the data



Step2 :

Data fields promote into header

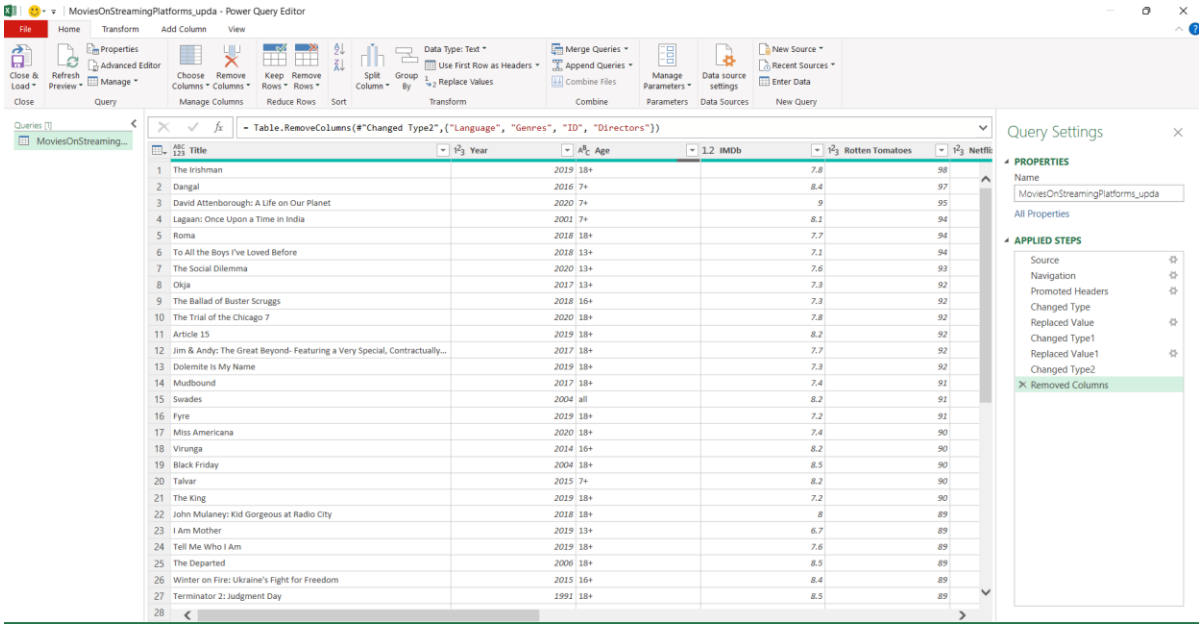


Step 3 :

Remove columns which are not required for the analysis

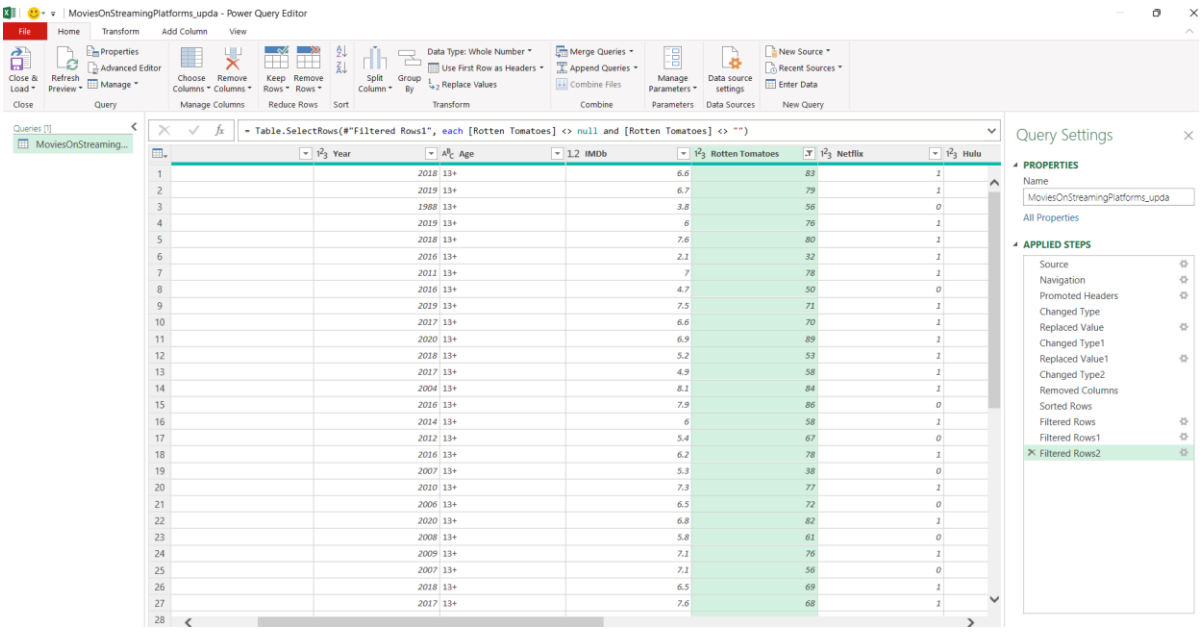
The columns here removed are :

- Language
- Genres
- Directors



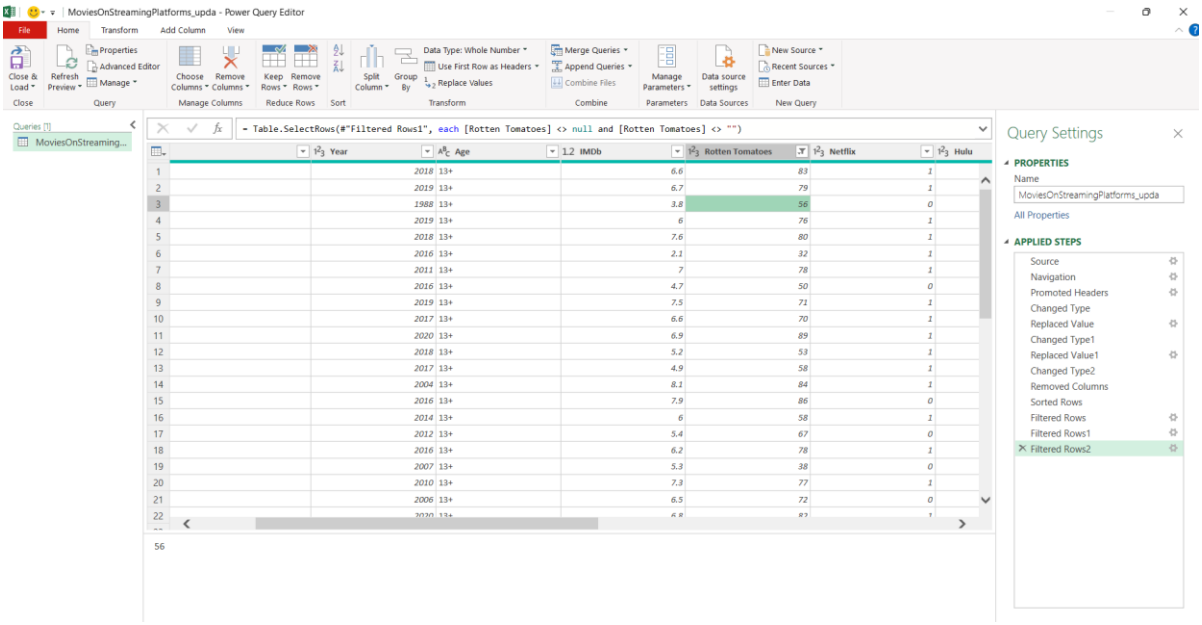
Step 4:

Remove null values by right clicking on the column and there we can remove empty fields and sort accordingly



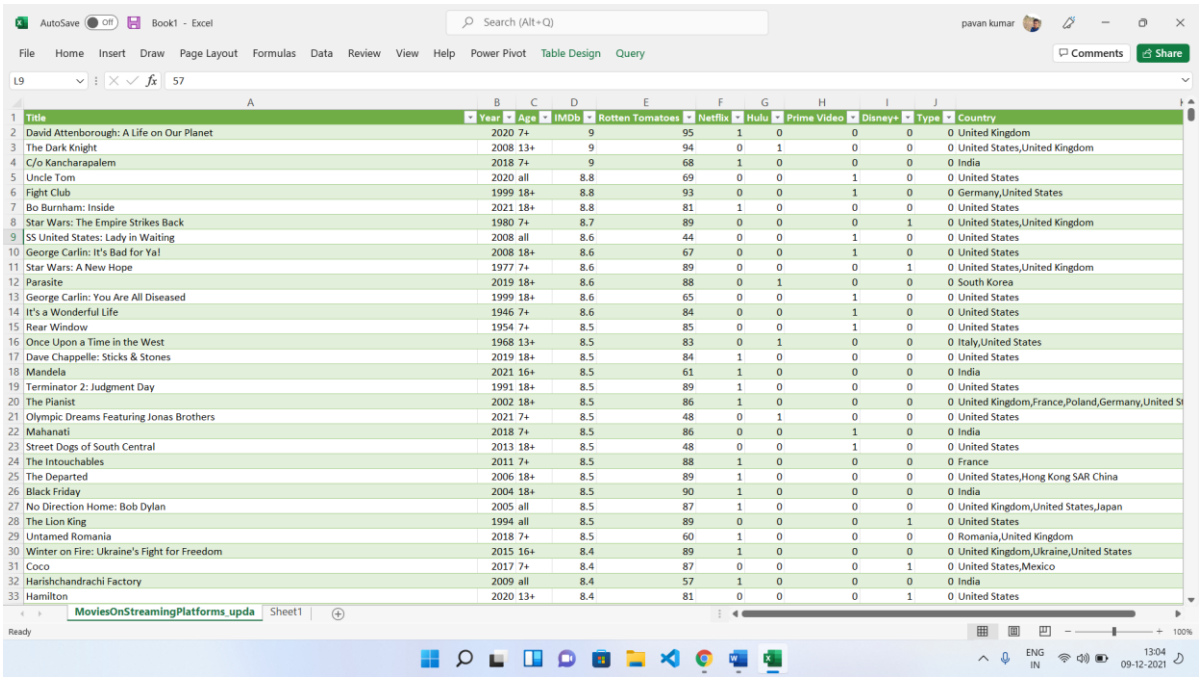
Step 5 :

Changed data types like rotten tomatoes are converted from text



Step 6 :

Now Close and load data in to excel

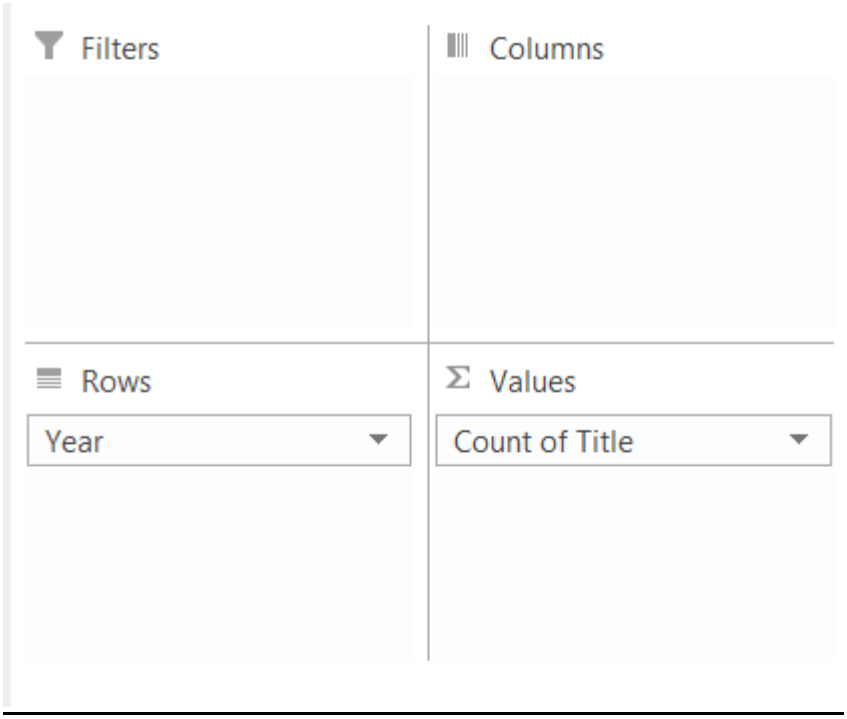


Title	Year	Age	IMDb	Rotten Tomatoes	Netflix	Hulu	Prime Video	Disney+	Type	Country
David Attenborough: A Life on Our Planet	2020	7+	9	95	1	0	0	0	0	United Kingdom
The Dark Knight	2008	13+	9	94	0	1	0	0	0	United States,United Kingdom
C/o Kancharapalem	2018	7+	9	68	1	0	0	0	0	India
Uncle Tom	2020	all	8.8	69	0	0	1	0	0	United States
Fight Club	1999	18+	8.8	93	0	0	1	0	0	Germany,United States
Bo Burnham: Inside	2021	18+	8.8	81	1	0	0	0	0	United States
Star Wars: The Empire Strikes Back	1980	7+	8.7	89	0	0	0	1	0	United States,United Kingdom
SS United States: Lady in Waiting	2008	all	8.6	44	0	0	1	0	0	United States
George Carlin: It's Bad for Ya!	2008	18+	8.6	67	0	0	1	0	0	United States
Star Wars: A New Hope	1977	7+	8.6	89	0	0	0	1	0	United States,United Kingdom
Parasite	2019	18+	8.6	88	0	1	0	0	0	South Korea
George Carlin: You Are All Diseased	1999	18+	8.6	65	0	0	1	0	0	United States
It's a Wonderful Life	1946	7+	8.6	84	0	0	1	0	0	United States
Rear Window	1954	7+	8.5	85	0	0	1	0	0	United States
Once Upon a Time in the West	1968	13+	8.5	83	0	1	0	0	0	Italy,United States
Dave Chappelle: Sticks & Stones	2019	18+	8.5	84	1	0	0	0	0	United States
Mandela	2021	16+	8.5	61	1	0	0	0	0	India
Terminator 2: Judgment Day	1991	18+	8.5	89	1	0	0	0	0	United States
The Pianist	2002	18+	8.5	86	1	0	0	0	0	United Kingdom,France,Poland,Germany,United St
Olympic Dreams Featuring Jonas Brothers	2021	7+	8.5	48	0	1	0	0	0	United States
Mahanati	2018	7+	8.5	86	0	0	1	0	0	India
Street Dogs of South Central	2013	18+	8.5	48	0	0	1	0	0	United States
The Intouchables	2011	7+	8.5	88	1	0	0	0	0	France
The Departed	2006	18+	8.5	89	1	0	0	0	0	United States,Hong Kong SAR China
Black Friday	2004	18+	8.5	90	1	0	0	0	0	India
No Direction Home: Bob Dylan	2005	all	8.5	87	1	0	0	0	0	United Kingdom,United States,Japan
The Lion King	1994	all	8.5	89	0	0	0	1	0	United States
Untamed Romania	2018	7+	8.5	60	1	0	0	0	0	Romania,United Kingdom
Winter on Fire: Ukraine's Fight for Freedom	2015	16+	8.4	89	1	0	0	0	0	United Kingdom,Ukraine,United States
Coco	2017	7+	8.4	87	0	0	0	1	0	United States,Mexico
Harishchandrachi Factory	2009	all	8.4	57	1	0	0	0	0	India
Hamilton	2020	13+	8.4	81	0	0	0	1	0	United States

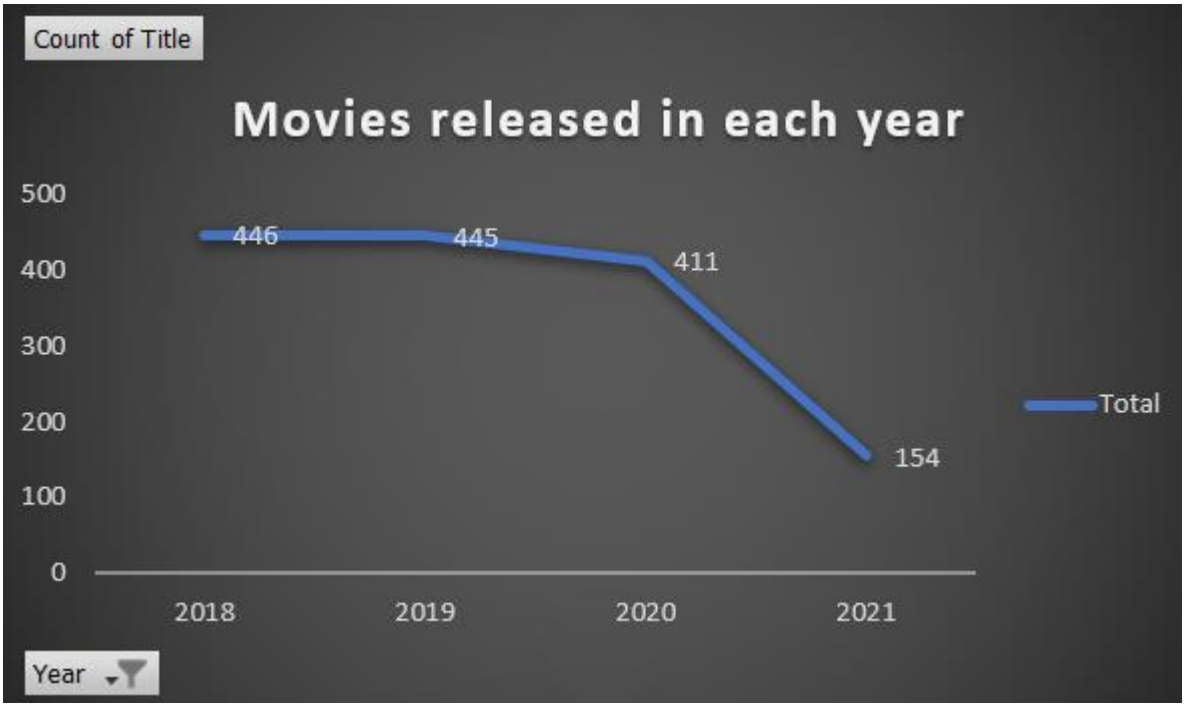
Data Analysis

1) Number of Movies released in each year

• firstly pivot table is created by clicking on insert tab in excel and select pivot table and then drag Years in to rows and darg titles to the values field so that it becomes count of the titles

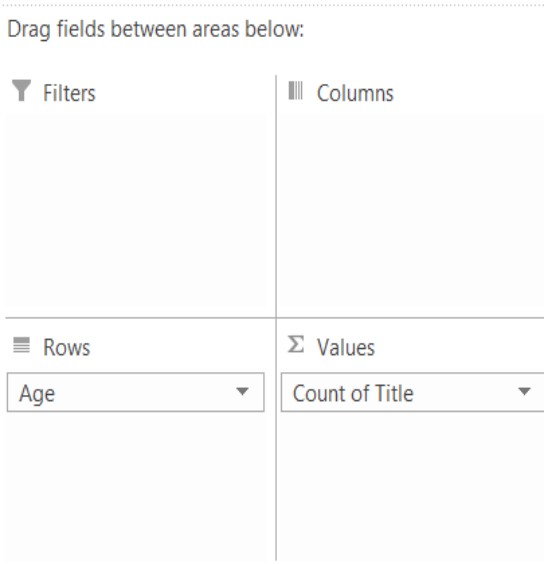


If we compare number of movies released in each year lets take 4 years that is 2018, 2019, 2020, 2021 using line chart the graph looks like this



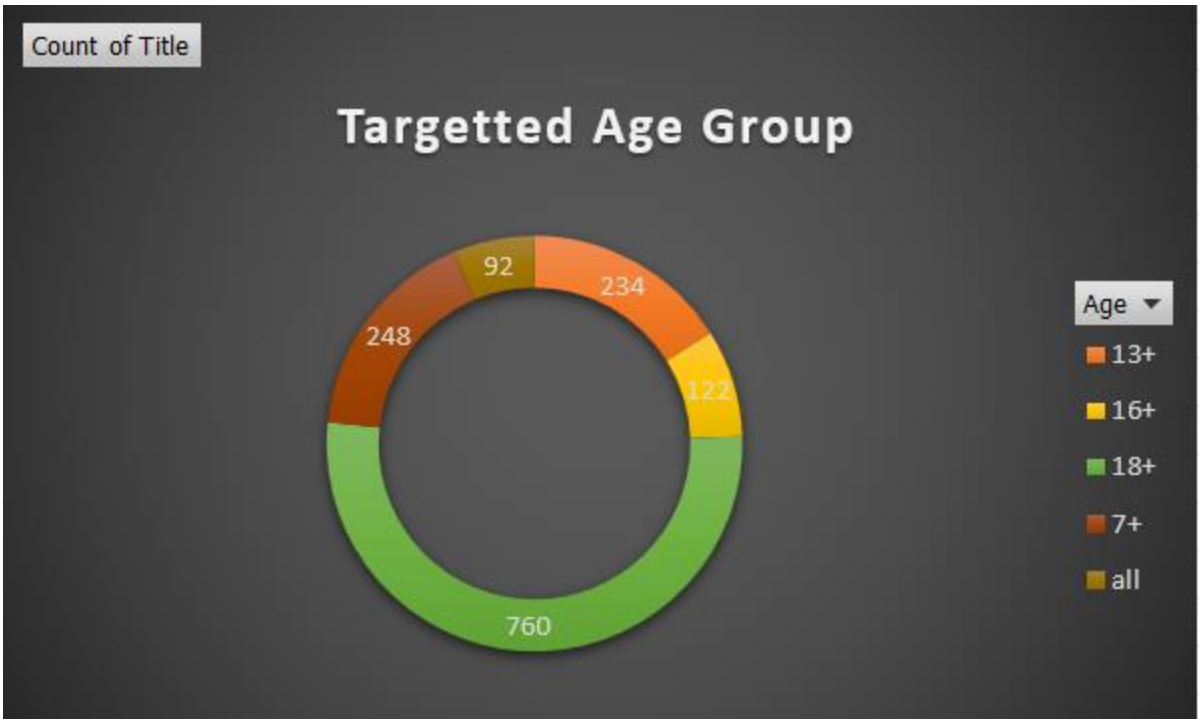
2) Targeted age group

In the pivot table Age column is dragged in to rows and count of titles in Values



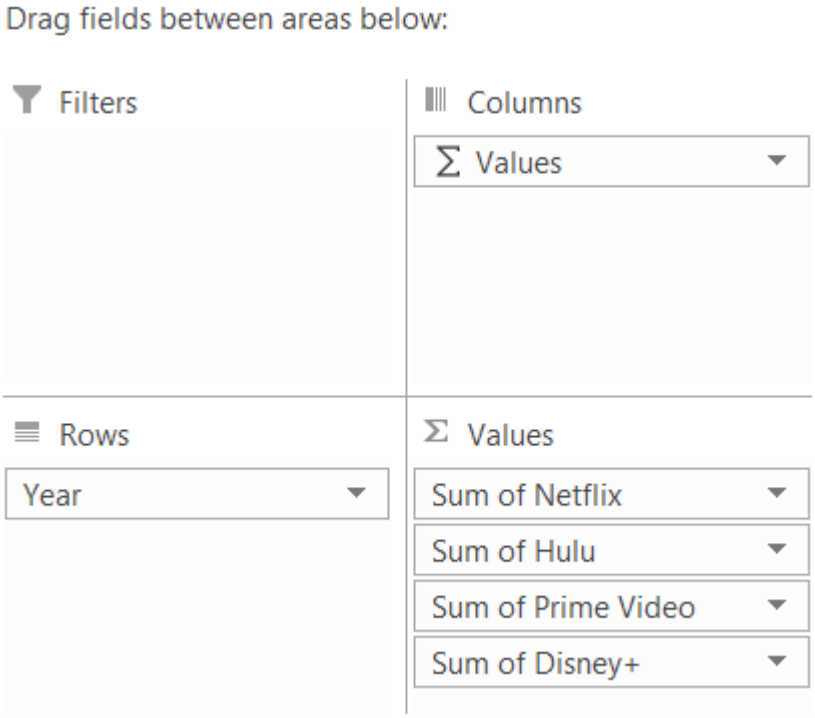
Row Labels	Count of Title
13+	234
16+	122
18+	760
7+	248
all	92
Grand Total	**1456**

Plotting this in pie chart looks like this



3) Number of movies released in Netflix, Hulu, prime, Disney+

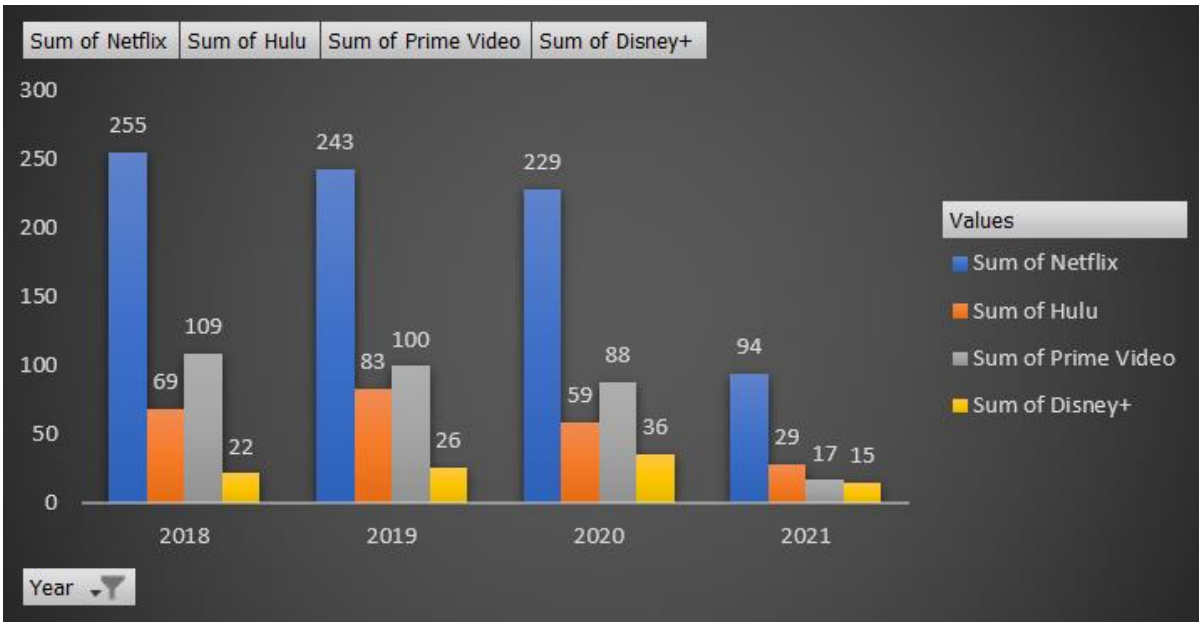
Here sum of Netflix, Hulu, Prime, Disney+ are dropped in to values and year is dropped in to rows in pivot table field



Pivot table for this analysis for four years looks like this

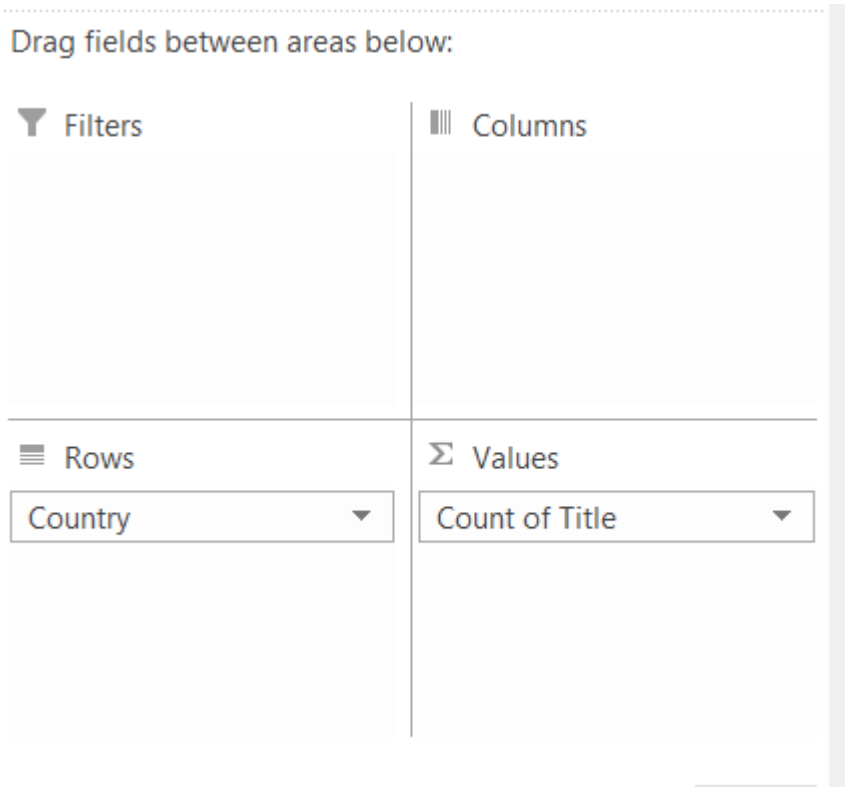
Row Labels	Sum of Netflix	Sum of Hulu	Sum of Prime Video	Sum of Disney+
2018	255	69	109	22
2019	243	83	100	26
2020	229	59	88	36
2021	94	29	17	15
Grand Total	**821**	**240**	**314**	**99**

And column chart for this analysis looks like this



4) Number of movies released per each country

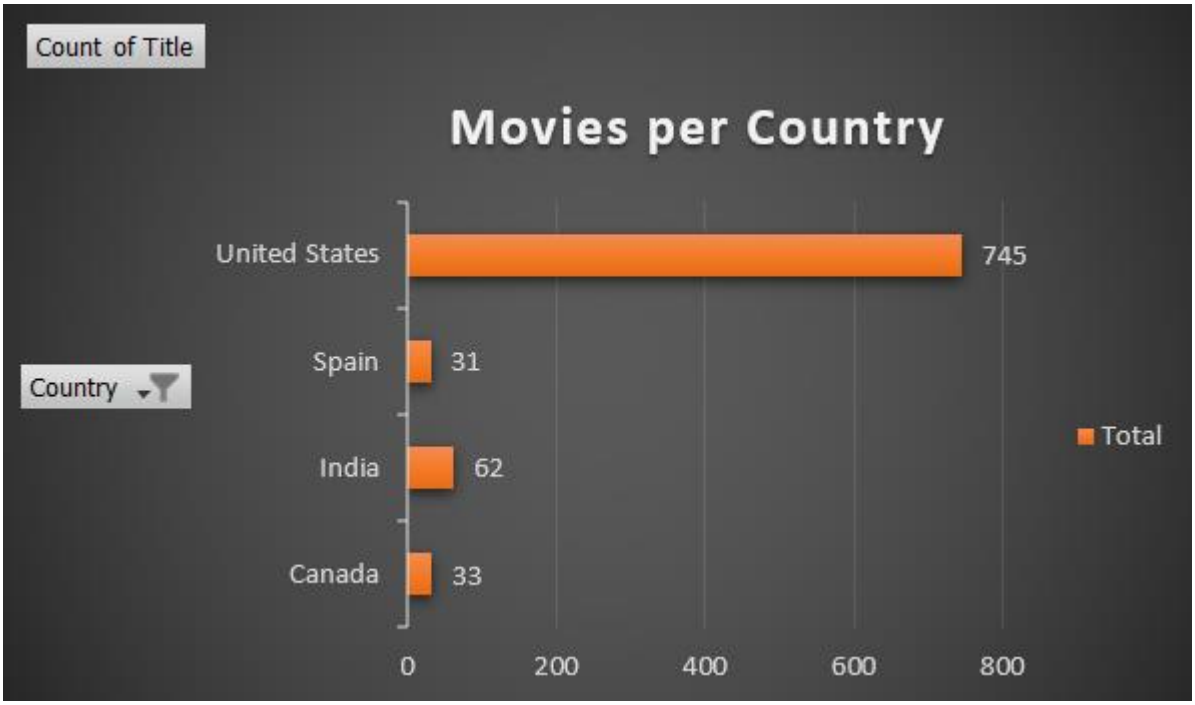
Here for this analysis countries is dropped into rows and count of the titles is dropped in to values



Pivot table for four countries that India, United States, Canada, Spain looks like this

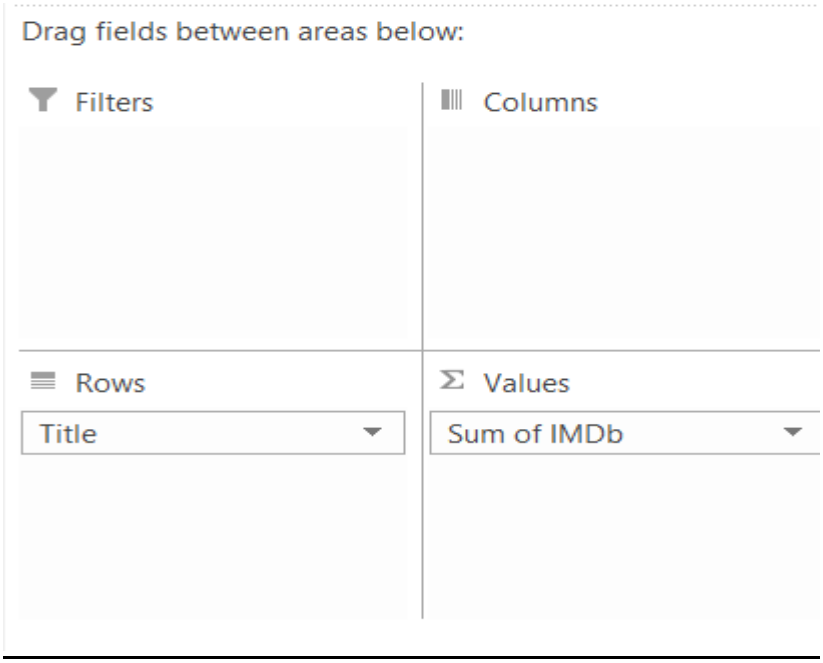
Row Labels	Count of Title
Canada	33
India	62
Spain	31
United States	745
Grand Total	**871**

Bar chart for this countries India, Spain, United States and Canada looks like this



5) Top 10 Movies According to IMDB

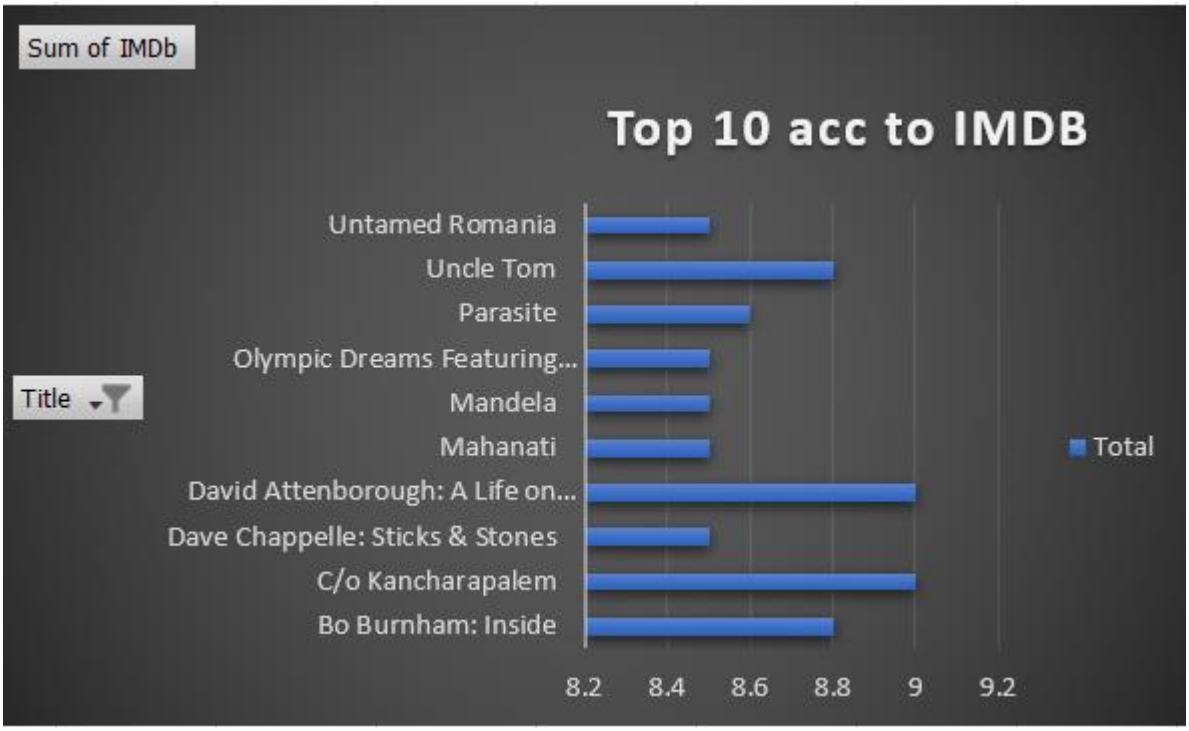
Titles are dropped in to rows and Sum of IMDB ratings are dropped in to values



In the row labels value filter is applied so that movies according to IMDB and pivot table looks like this

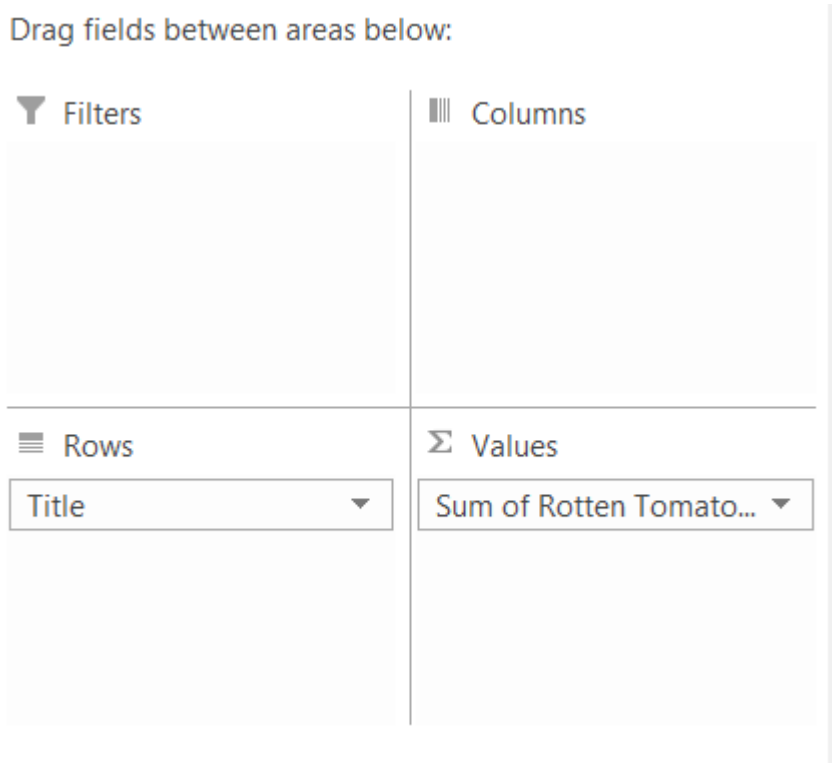
Row Labels	Sum of IMDb
Bo Burnham: Inside	8.8
C/o Kancharapalem	9
Dave Chappelle: Sticks & Stones	8.5
David Attenborough: A Life on Our Planet	9
Mahanati	8.5
Mandela	8.5
Olympic Dreams Featuring Jonas Brothers	8.5
Parasite	8.6
Uncle Tom	8.8
Untamed Romania	8.5
Grand Total	**86.7**

Top 10 Movies in Bar chart



6) Top 10 Movies According To Rotten Tomatoes

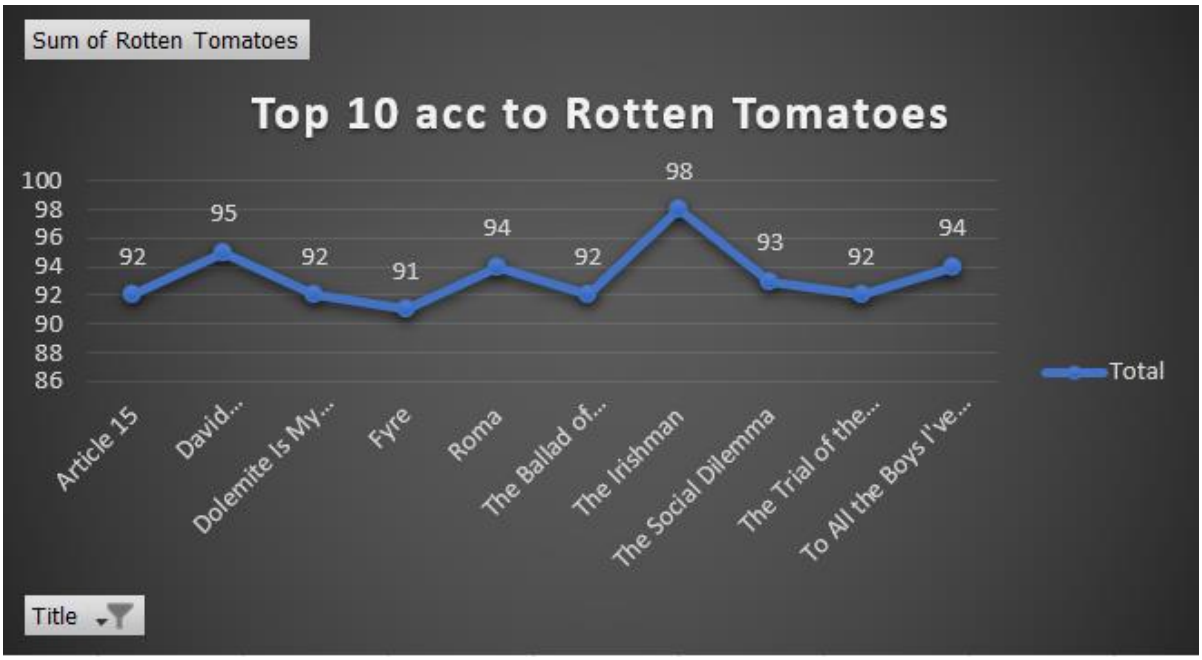
Titles are dropped in to rows and Sum of Rotten Tomatoes ratings are dropped in to values



In the row labels value filter is applied so that top 10movies according to Rotten Tomatoes and pivot table looks like this

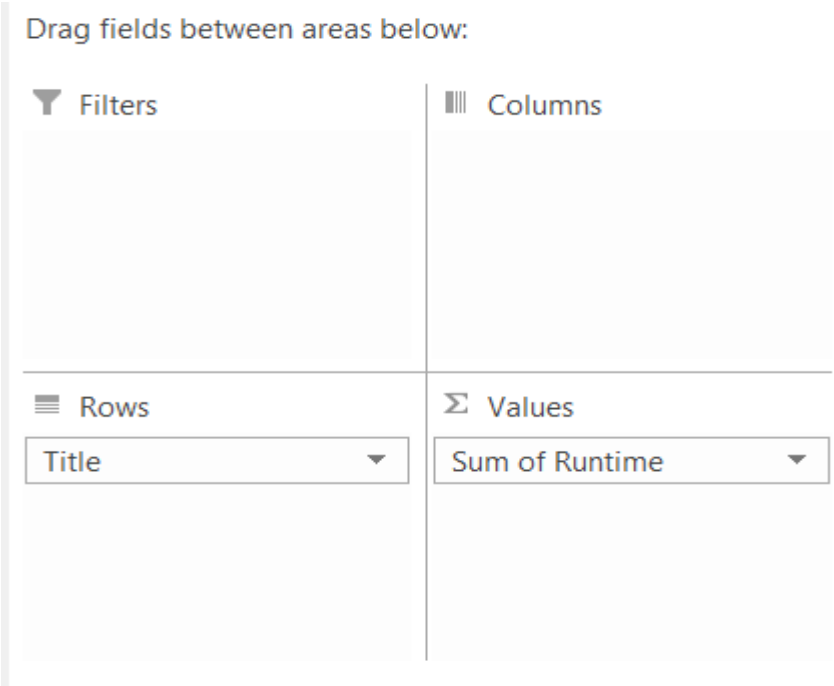
Row Labels	Sum of Rotten Tomatoes
Article 15	92
David Attenborough: A Life on Our Planet	95
Dolemite Is My Name	92
Fyre	91
Roma	94
The Ballad of Buster Scruggs	92
The Irishman	98
The Social Dilemma	93
The Trial of the Chicago 7	92
To All the Boys I've Loved Before	94
Grand Total	**933**

Top 10 Movies According to Rotten Tomatoes Line chart



7) Longest Run Time Movies

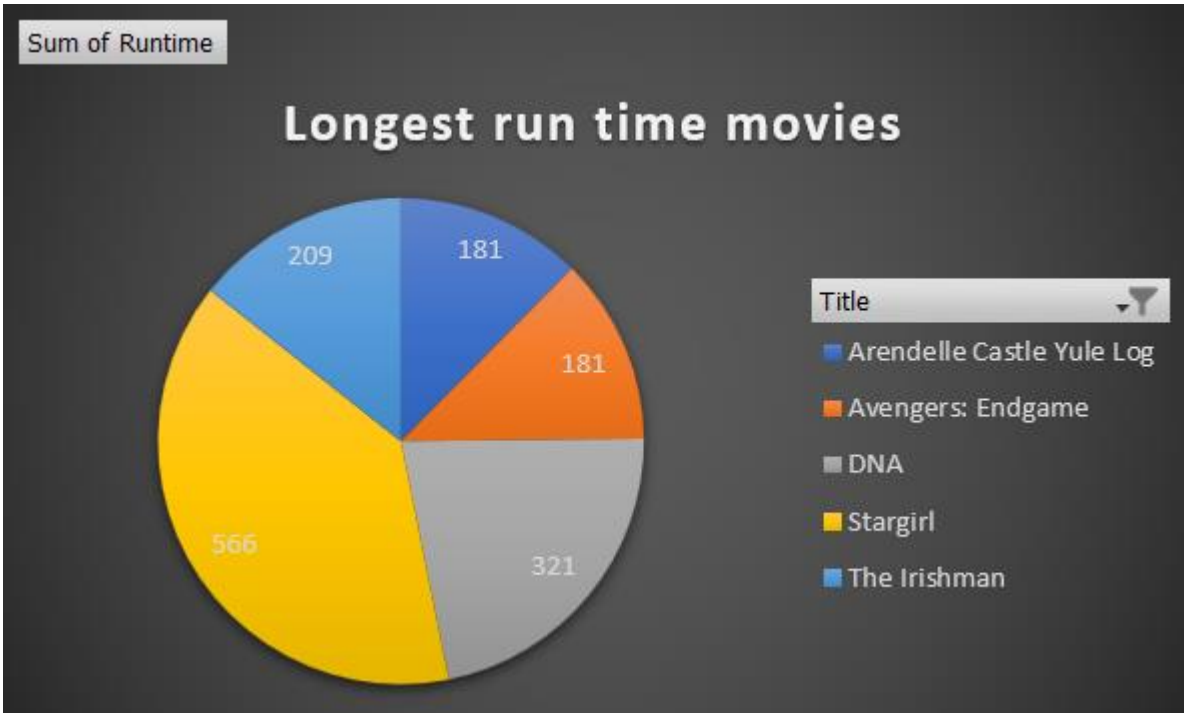
For this analysis Titles are dropped in to rows and sum of Run time is dropped in to values



In the row labels value filter is applied so that top 5 movies With the longest run time in pivot table looks like this

Row Labels	Sum of Runtime
Arendelle Castle Yule Log	181
Avengers: Endgame	181
DNA	321
Stargirl	566
The Irishman	209
Grand Total	**1458**

When Pie Chart used For Analysis To find top 5 movies with Longest Run time looks like this



Dashboard :

Firstly in the page layout tab view gridlines box is un checked so that grid lines will get disappeared

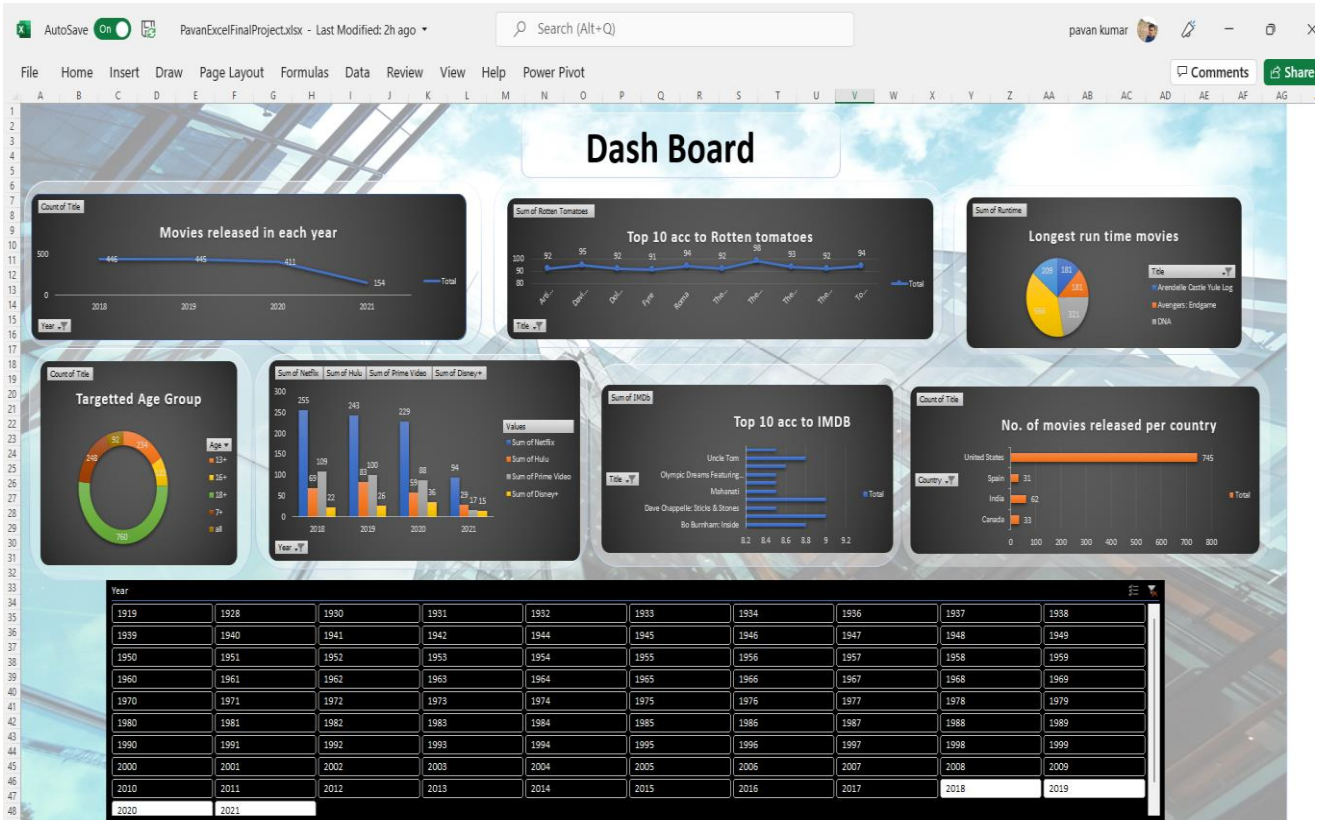
Secondly in the illustrations the rectangle shape is drawn and from the page layout background image is added and transparency set accordingly and adjusted the image

Thirdly for all the 7 charts the 7 rectangular shapes with rounded edges are drawn with adjusted transparency and No line is selected so that the out line will not get appear so that the shapes looks more clean

Charts from all the other sheets copied and pasted in the shapes

Lastly by clicking on any graph slicer is included and then in the slicer tab in report connectivity tab and check for all the pivot tables so that all the charts of all the pivot tables are connected to one slicer and for the slicer design first create a duplicate theme in any of the theme and then after creating duplicate right click on that theme and all the settings were modified

Image of the Final Dash Board



BIBLIOGRAPHY

Source of Data Set : https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

Data Management :https://support.microsoft.com/en-us/office/using-access-or-excel-to-manage-your-data-09576147-47d1-4c6f-9312-e825227fcaea

THE END