Summary: Visualizing Netflix Dataset

SESSION OVERVIEW:

By the end of this session, students will be able to:

- Analyse and visualise a real-life dataset and generate insights out of it.
- Get answers to questions around a dataset through Power BI visualisations.

KEY TOPICS AND EXAMPLES:

Note: For this session, we will be using this Dataset.
In this session we will be visualising the Netflix dataset in Power BI.

1. Understanding the Dataset

The dataset needs to be understood and explored first. This step can be done in either Excel or Power BI. We will scrutinise the useful fields to understand the meaning and data distribution of each.

- **id**: A unique count of this ID would be useful while calculating and understanding trends. Example: The number of movies released in 2020 would simply be a unique count of the **ID** field.
- **release_year**: At the outset, it seems that it includes movies from 1953 to 2022, i.e. 70 years of data. But on closer inspection, we notice that some years are missing e.g. 1955, 1957.
- **title**: The dataset becomes very interesting as many of the titles can be recognised.
- **type**: If we use the FILTER option in Excel to see the different TYPE values that are there, we will see two: MOVIE and SHOW.
- **description**: Although for our analysis, description may not be very important but we can extract keywords out of it and use machine learning to extract information like the genre of the movie/show.
- **age_certification**: Something to note here is that in many cases age certification is missing. We may want to handle missing values separately in some cases, but here since there is no effective way to substitute this missing information, we will leave it as is for our analysis.
- **runtime**: Although it is not mentioned, it is clear that runtime is in **minutes**. One should see the range of values, visualise the distribution and if required, handle outliers in this field.
- **imdb_score**: This is the IMDb rating of the movie/show. This tells us how well the movie was rated by the voters.
- **imdb_votes**: This is the number of people who voted on IMDb for a movie or show. It is important to recognise that this leads to a lot of subjectivity because if very few people have voted for a movie/show its imdb_score will not be very reliable.

The dataset has 5283 records and this data includes both movies and shows.

2. Questions that PowerBI will Answer

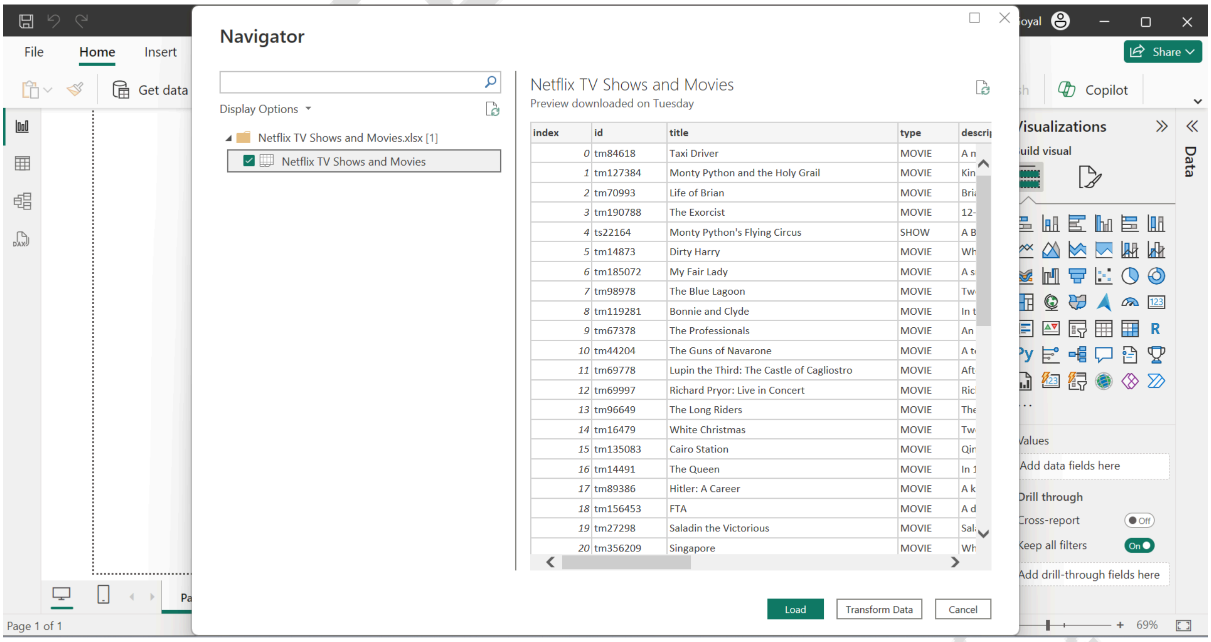
At this point it is important to recognise some of the questions that we can answer through this dataset. In organisations, while solving business problems, some of these questions will be given to you by the business stakeholder.

1. Which was the best movie and TV show overall in the last 50 years?
2. How many movies do we have in our dataset across the last few years? Do we have more representation of movies from the last 20 years or is the dataset free from any such skewness?
3. On average, how has the IMDb score been trending over the last 50 years? Has it been deteriorating or improving?
4. Have more people started voting for movies/shows on IMDb over the last 50 years?
5. On average, how has the runtime changed over the last 50 years?
6. How does age certification of a movie affect its rating?

We will cover the answers to all these questions using this dataset in Power BI.

3. Data Cleaning and Manipulation in Power Query Editor

We need to import the dataset in Power BI and click on Transform.



Netflix TV Shows and Movies
Preview downloaded on Tuesday

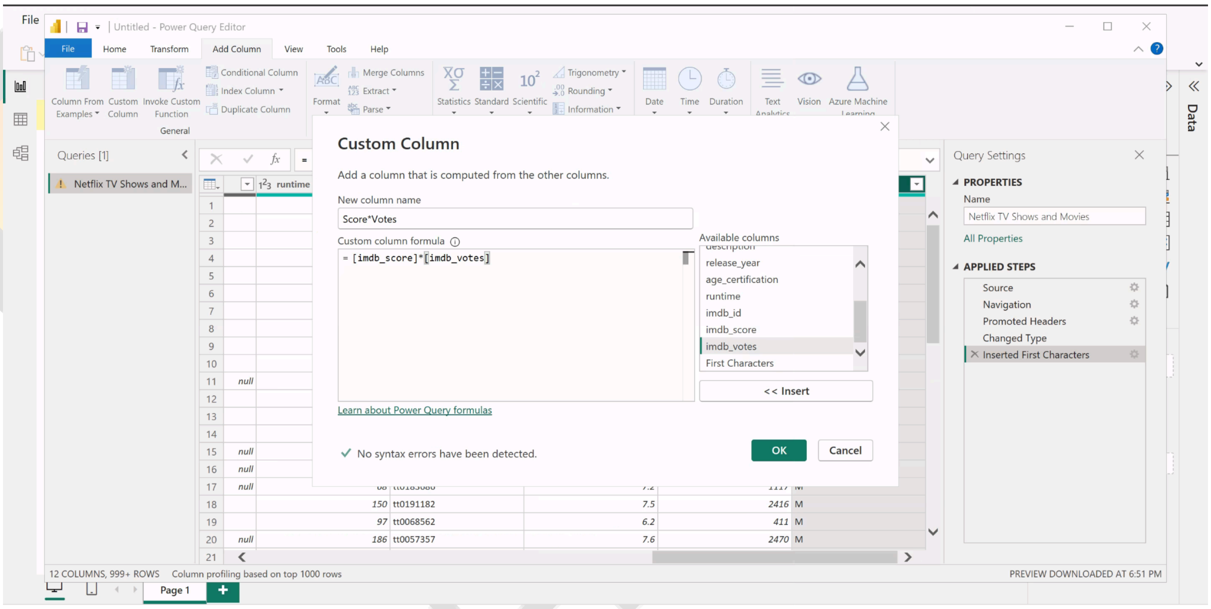
index	id	title	type	descri
0	tm84618	Taxi Driver	MOVIE	A n
1	tm127384	Monty Python and the Holy Grail	MOVIE	Kin
2	tm70993	Life of Brian	MOVIE	Bri
3	tm190788	The Exorcist	MOVIE	12-
4	ts22164	Monty Python's Flying Circus	SHOW	A B
5	tm14873	Dirty Harry	MOVIE	Wh
6	tm185072	My Fair Lady	MOVIE	A s
7	tm98978	The Blue Lagoon	MOVIE	Tw
8	tm119281	Bonnie and Clyde	MOVIE	In t
9	tm67378	The Professionals	MOVIE	An
10	tm44204	The Guns of Navarone	MOVIE	A t
11	tm69778	Lupin the Third: The Castle of Cagliostro	MOVIE	Aft
12	tm69997	Richard Pryor: Live in Concert	MOVIE	Ric
13	tm96649	The Long Riders	MOVIE	The
14	tm16479	White Christmas	MOVIE	Tw
15	tm135083	Cairo Station	MOVIE	Qir
16	tm14491	The Queen	MOVIE	In 1
17	tm89386	Hitler: A Career	MOVIE	A k
18	tm156453	FTA	MOVIE	A d
19	tm27298	Saladin the Victorious	MOVIE	Sal
20	tm356209	Singapore	MOVIE	Wh

To make more sense out of our Netflix Dataset, one manipulation we can do is to create a custom column that is a combination of both IMDb score (imdb_score) and number of votes (imdb_votes). Score alone may not be sufficient to conclude anything about a movie/show.

Example: What if only one person voted for the movie and the rating given by that person is a 9.5. Does this mean that the movie is better than most of the movies? Maybe not.

We would want to prefer watching a movie that is rated highly by many people. We will hence definitely prefer one that has a 9 rating but has been voted by 50000 people. Hence, we should not see either IMDb score or IMDb votes in isolation. It may be more useful for us to combine both columns and one meaningful way to do that is to multiply scores and votes. This is where a custom column becomes useful.

The image below demonstrates this formula.



Custom Column

Add a column that is computed from the other columns.

New column name

Score*Votes

Custom column formula

```
= [imdb_score]*[imdb_votes]
```

Available columns

description
release_year
age_certification
runtime
imdb_id
imdb_score
imdb_votes
First Characters

<< Insert

Learn about Power Query formulas

✓ No syntax errors have been detected.

OK Cancel

Since we want to use both **scores** and **votes** to evaluate a movie, we create a new field for [imdb_score] * [imdb_votes].

Alternative Approach: While doing any analysis that requires rating/score, we can consider ratings where **votes are more than 1000**. This ensures that only movies that are voted for by a sufficient number of people are considered.

Further, we can drop the **index** and **imdb_id** columns as they may not be useful for our analysis and will not be required.

A closer look at imdb_votes reveals that there are movies/shows where IMDb votes are absent. While we cannot remove these rows since they have other valuable data, we can ignore these rows when analysing votes.

After doing the above cleaning/manipulation, click on close and apply to load the data.

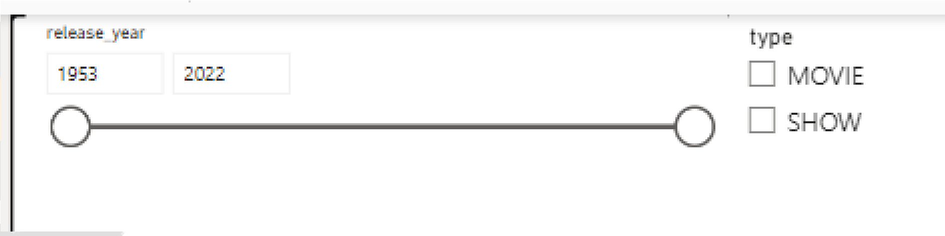
4. Loading Data and Creating a Model

After cleaning and manipulation is done, we can move to creating a model in Power BI. Since we are using only one dataset, it is not required to create a model. If required at this point, we can also create DAX measures, calculated columns and calculated tables.

5. Visualisations

It is a good practice to start with selecting a theme.

Next, we can create slicers. In this case we will create two slicers: one for type and the other for release year.



release_year

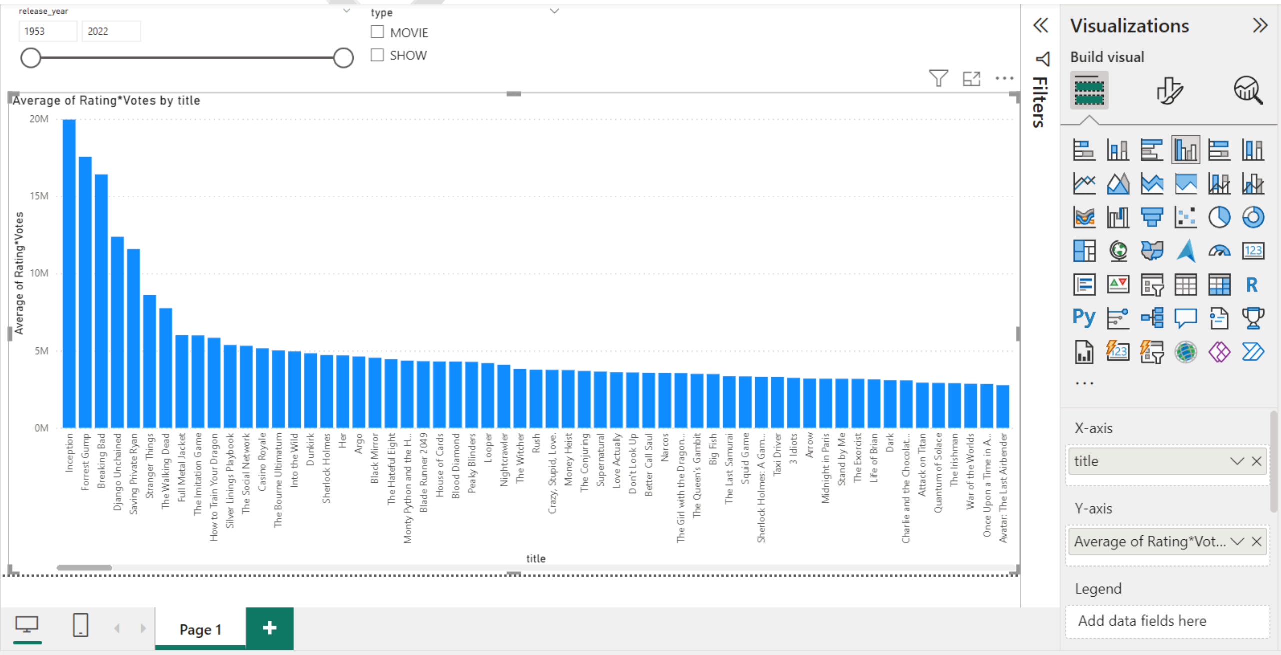
1953 2022

type

□ MOVIE

□ SHOW

1. **Which was the best movie and TV show overall in the last 50 years?**
   For answering this question, we will create a column chart for movie/show title and (Rating x Votes). This will immediately tell us the best movie/show in the dataset. It has taken both rating/score and votes into account and hence is a good way to answer the question about the best movie/show.



release_year

1953 2022

type

□ MOVIE

□ SHOW

X-axis

title

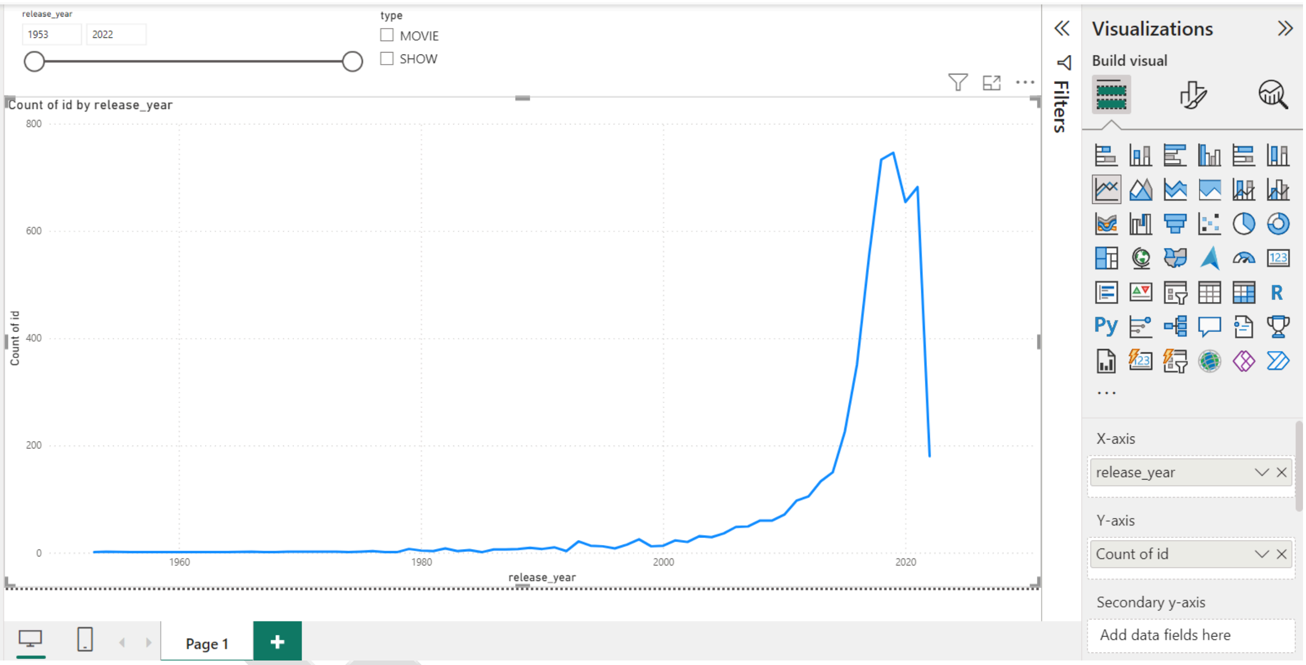
Y-axis

Average of Rating*Vot...

Legend

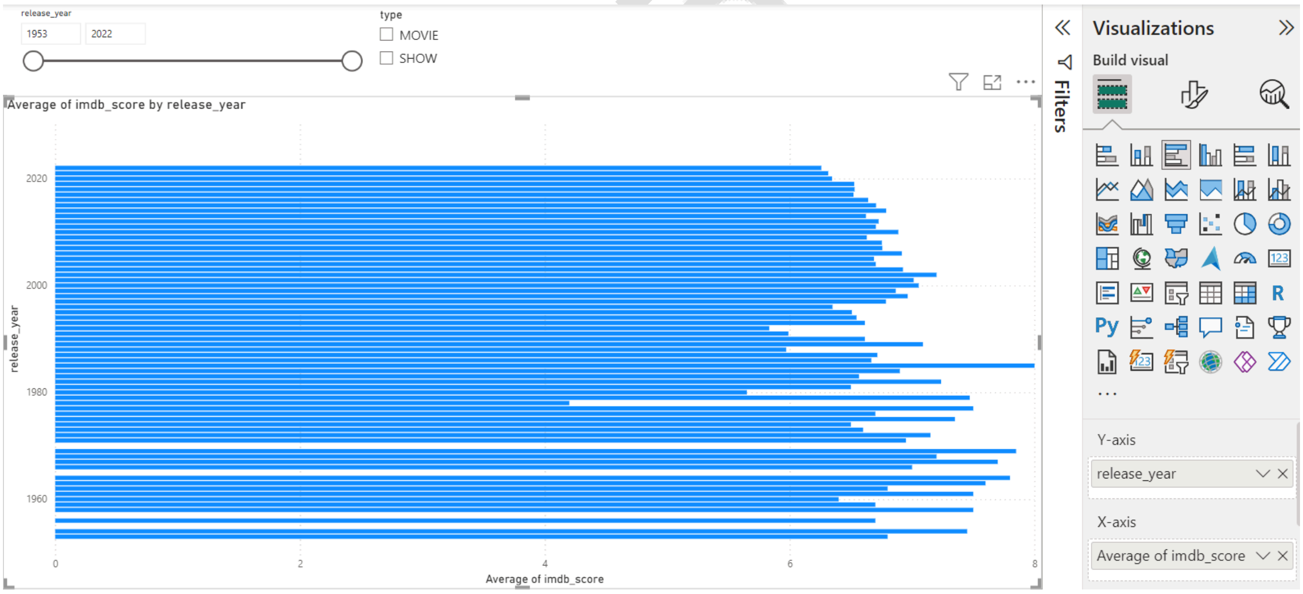
Add data fields here

2. **How many movies do we have in our dataset across the last few years?**
   For this, we can create a line chart with release year on the horizontal axis and distinct count of IDs on the vertical axis as shown below.

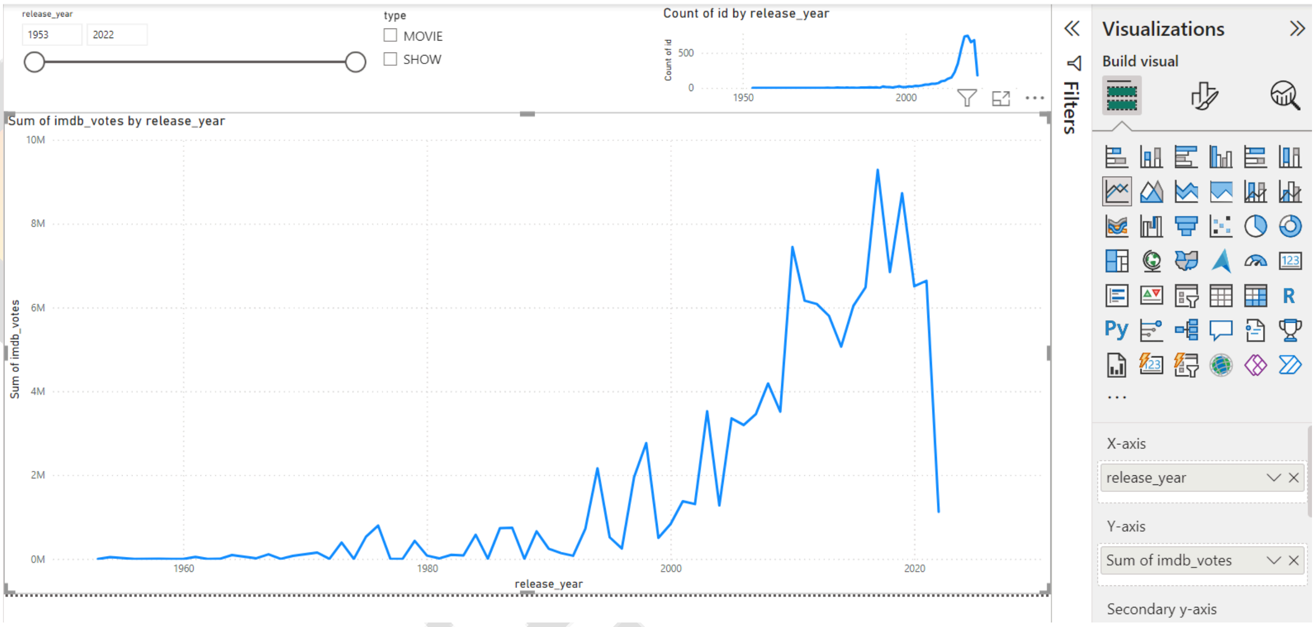


As we can see from the above chart the dataset is skewed towards the right - we have a high number of movies from after the 2010 period as compared to that before 2000. While making any conclusion, this needs to be kept in mind.

3. **Across the last 50 years, how has the average IMDb score varied?**
For this we will use a bar chart with release year on the y-axis and average IMDb score on the x-axis.



4. **Have more people started voting for movies/shows on IMDb over the last 50 years?**
   We will use a line chart for this, with release year on x-axis and total IMDb votes on y-axis.

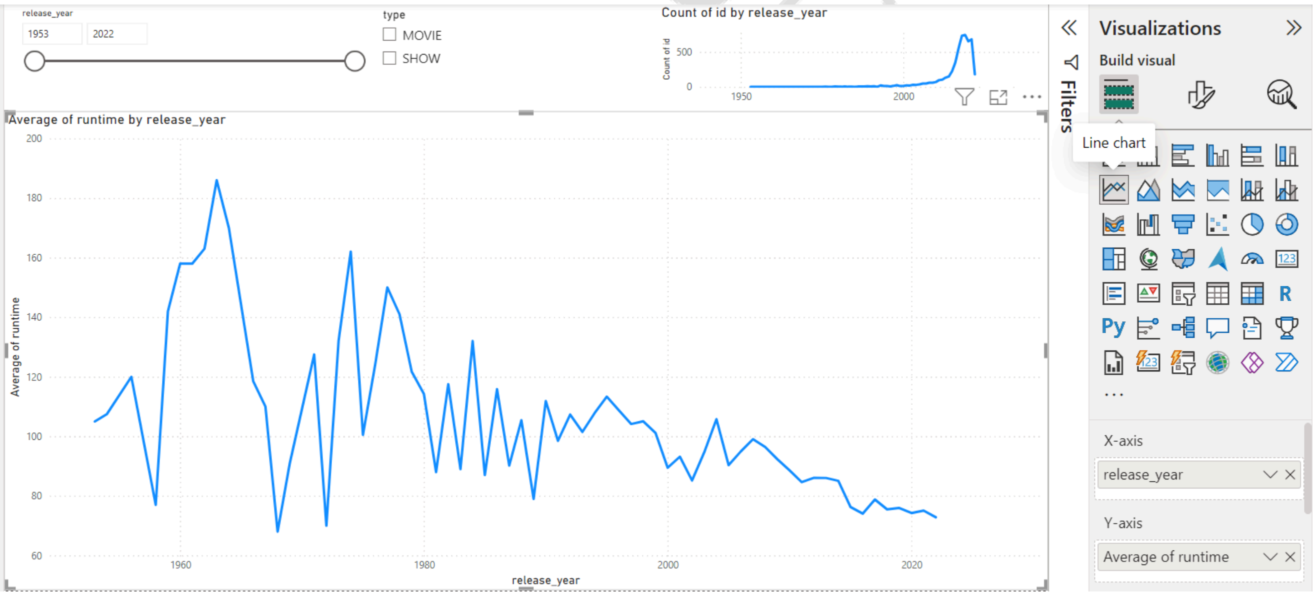


X-axis
release_year

Y-axis
Sum of imdb_votes

Note that the reason for the above trend is also the high number of movies in the post-2010 period. Hence, we should not conclude any useful insights from this chart and it may not be included in the final report.

5. **On average, how has the runtime changed over the last 50 years?**
   For this, we can plot the average runtime on a line chart with release year.



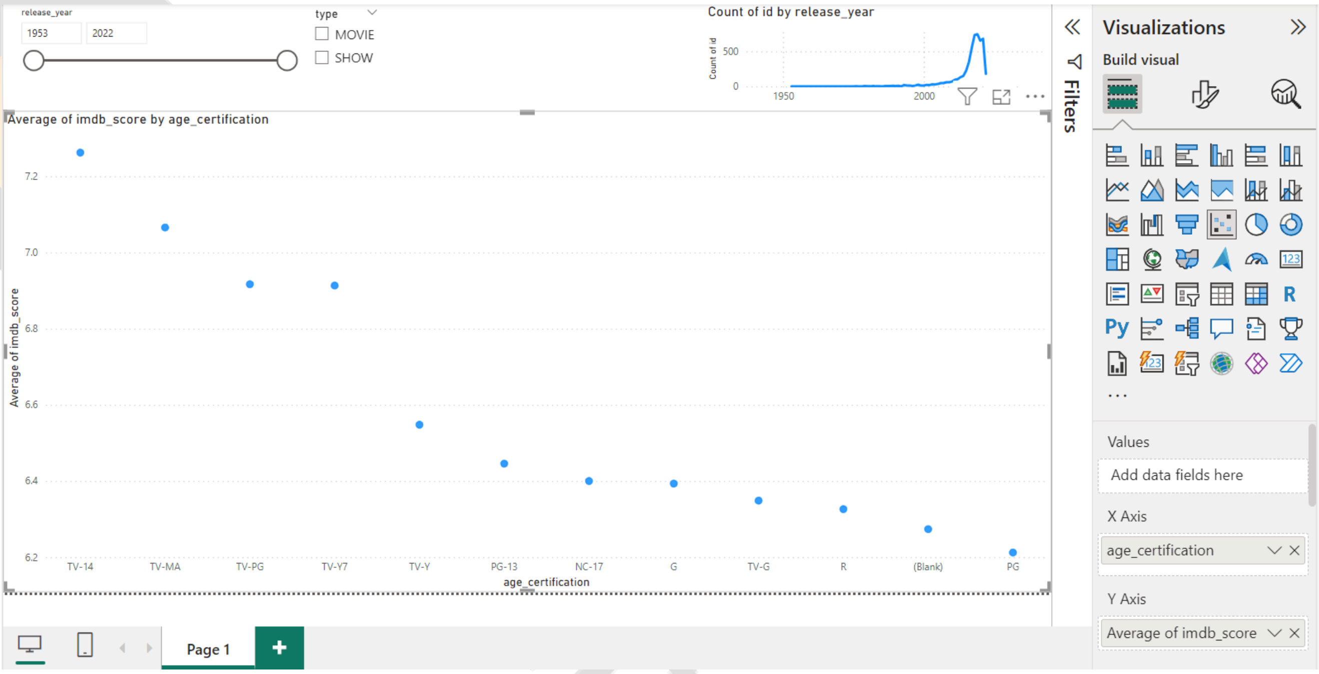
X-axis
release_year

Y-axis
Average of runtime

As we can see from this chart, the trend has become more consistent and the runtime has broadly decreased. It is important to understand that part of the reason for this trend may be the few datapoints we have for the early years: 1950s, 1960s, 1970s, 1980s and 1990s.

6. How does age certification of a movie affect its rating?

For answering this question, we will plot the age certification along with the average rating on a scatter plot.

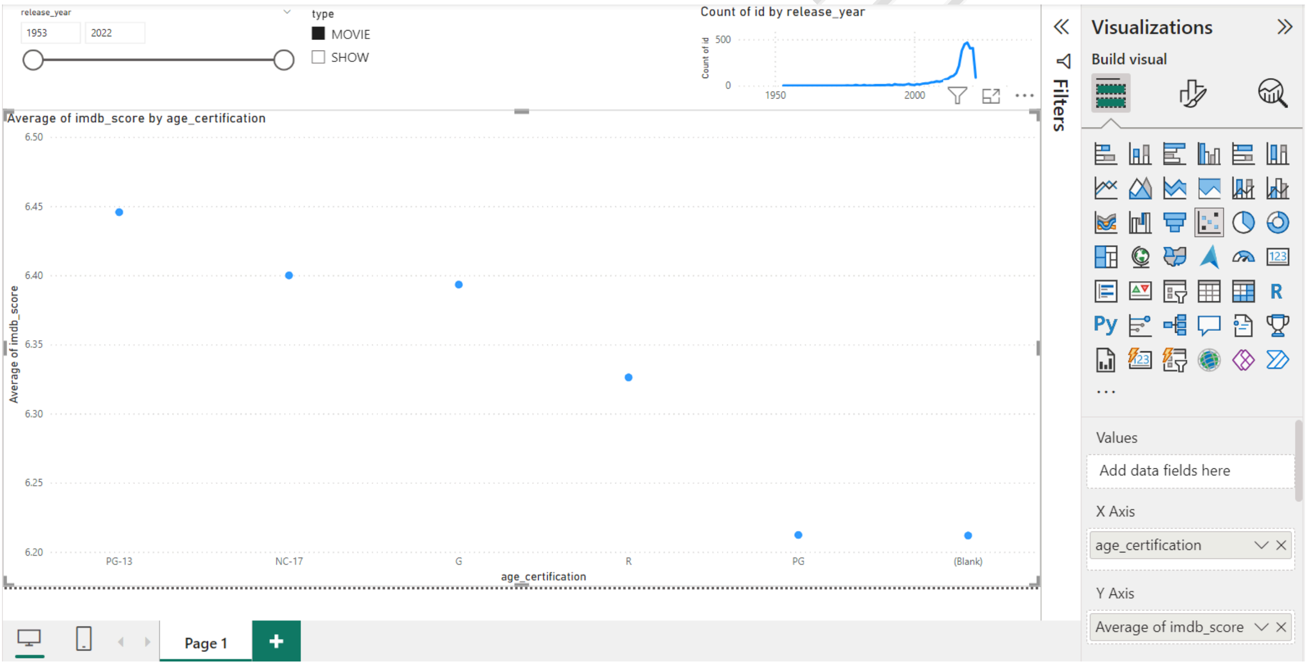


X Axis: age_certification

Y Axis: Average of imdb_score

We have three insights from this:

- On average, TV shows have higher ratings as compared to movies.
- On average, within TV shows, TV-14 gets the highest rating.
- On average, within Movies, PG-13 movies have the highest rating. This can also be seen from the image below.



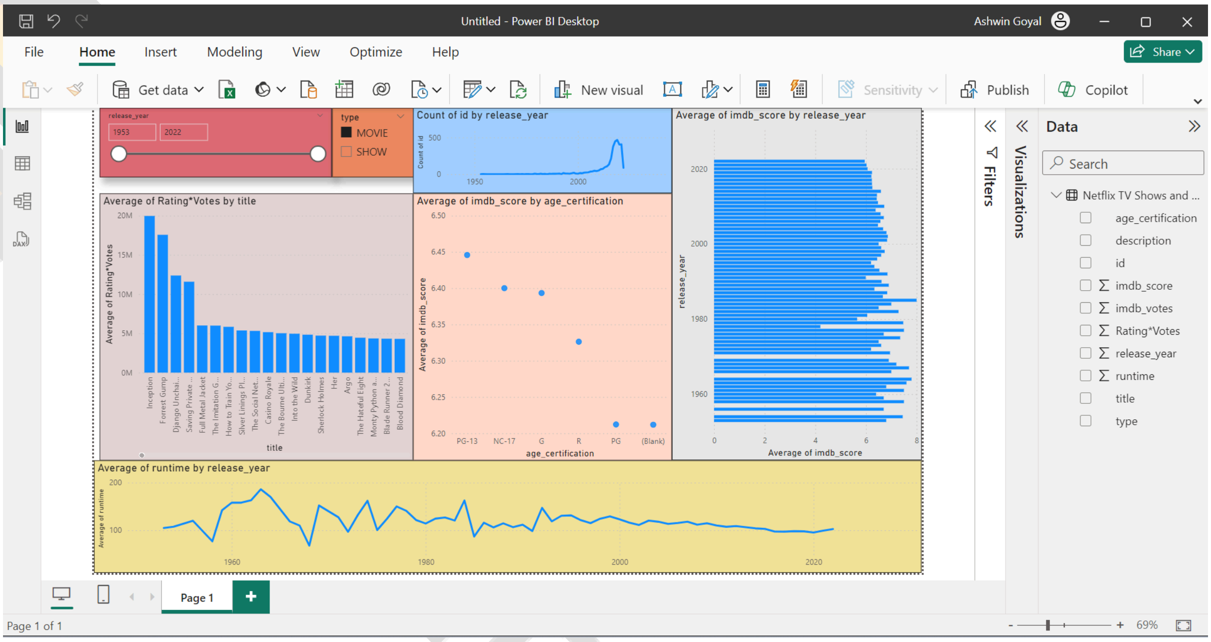
type: ■ MOVIE □ SHOW

X Axis: age_certification

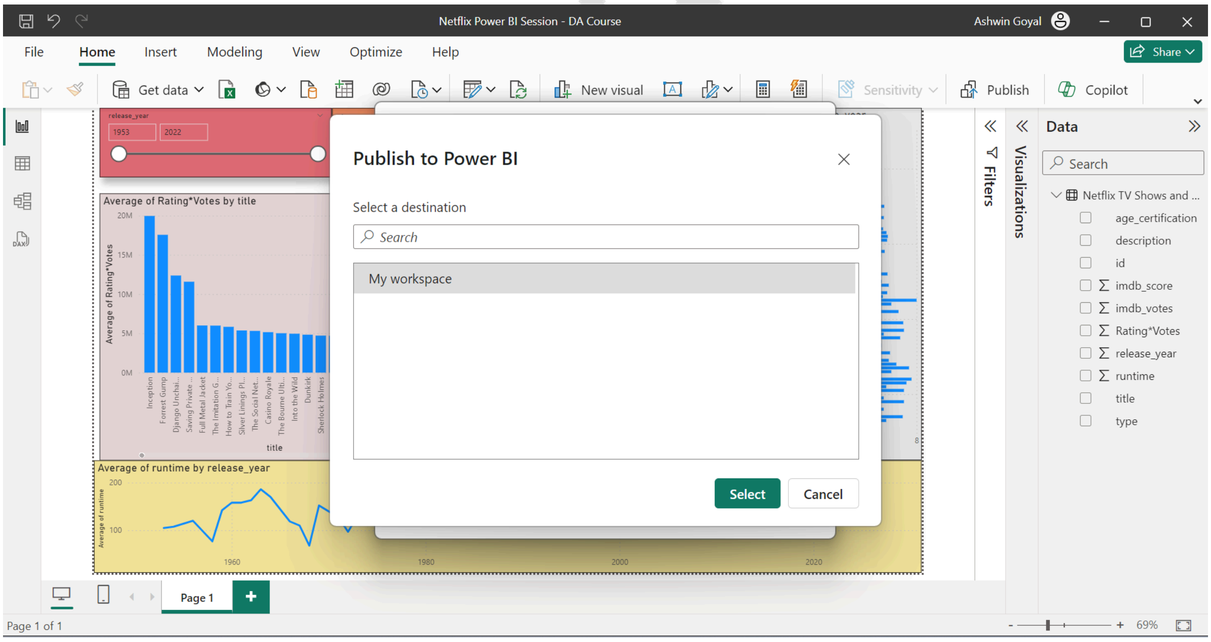
Y Axis: Average of imdb_score

6. Data Storytelling

All the questions that we set out to answer have been answered through visualisations. Now we need to make the visuals and the report visually appealing. For this we will use the formatting options for each visual and will arrange the visuals to create a story. If required we will use bookmarks and drill through functionalities. We can create multiple pages and use them to create a story around the Netflix data.



We can publish this dataset to Power BI server in a workspace as shown here.


Netflix Power BI Session - DA Course
Ashwin Goyal
File
Home
Insert
Modeling
View
Optimize
Help
Share
Get data
New visual
Sensitivity
Publish
Copilot
Publish to Power BI
Select a destination
Search
My workspace
Select
Cancel
Filters
Visualizations
Data
Search
Netflix TV Shows and ...
age_certification
description
id
imdb_score
imdb_votes
Rating*Votes
release_year
runtime
title
type
Page 1
Page 1 of 1
69%