PROJECT 1

CINEMATIC TRENDS: EXPLORING MOVIE METADATA AND VIEWER PREFERENCES

Team Members:

Sashidhar Chary Viswanathula (11714360)

Dimpu Nithish Karanam (11823599)

Vidya Gangasani (11702121)

Problem Statement:

The movie industry produces a huge amount of data, from movie genres and release dates, ratings and popularity metrics. But it can be challenging to understand what truly drives a movie's success or failure. This project focuses on uncovering insights into viewer preferences, genre popularity, trends in movie ratings over time, and Profits earned by a movie.

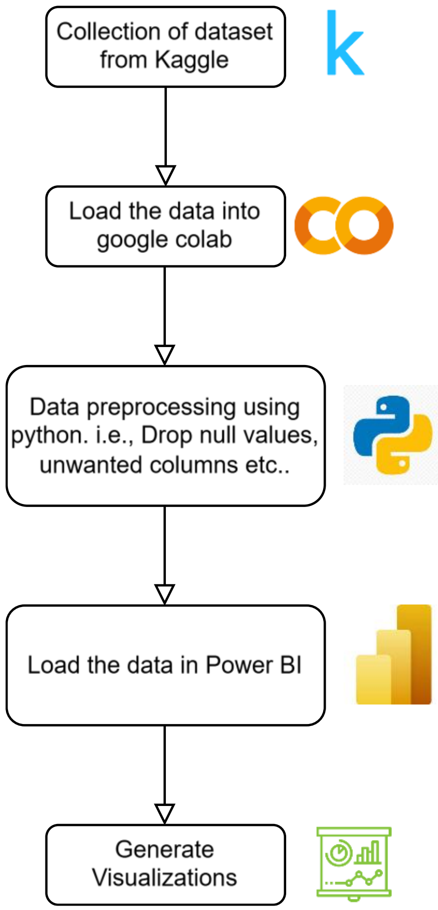
By analyzing the two datasets of movie metadata and ratings, we plan to answer questions like:

- Do certain genres consistently receive higher ratings?
- How does a movie's release year or runtime influence audience perception?
- Are there patterns in viewer behavior that correlate with popularity or average rating?
- What is the average budget of a genre, and average profits earned?

We are focusing on communicating these insights through interactive and engaging dashboards that reveal how people engage with films.

Workflow Diagram:

This is the Workflow diagram for the project 1, We started off by collecting the dataset from Kaggle and loading it to the Google collab, after that preprocessed the dataset using Python and loaded the preprocessed dataset into power BI to generate meaningful insights and visualizations.



Data Abstraction:

For this project, we are using datasets from Kaggle, which are publicly available and have rich information about the movies and their ratings. We've used a structured dataset, but it has complex columns which have a JSON type in it. We processed the JSON type columns to get the data we needed from those columns.

movies_metadata.csv (From Kaggle – TMDB dataset):

Contains information about movies like titles, genres, budget, revenue, runtime, popularity, release dates, production companies, and more. This will help us analyze movie characteristics and metadata.

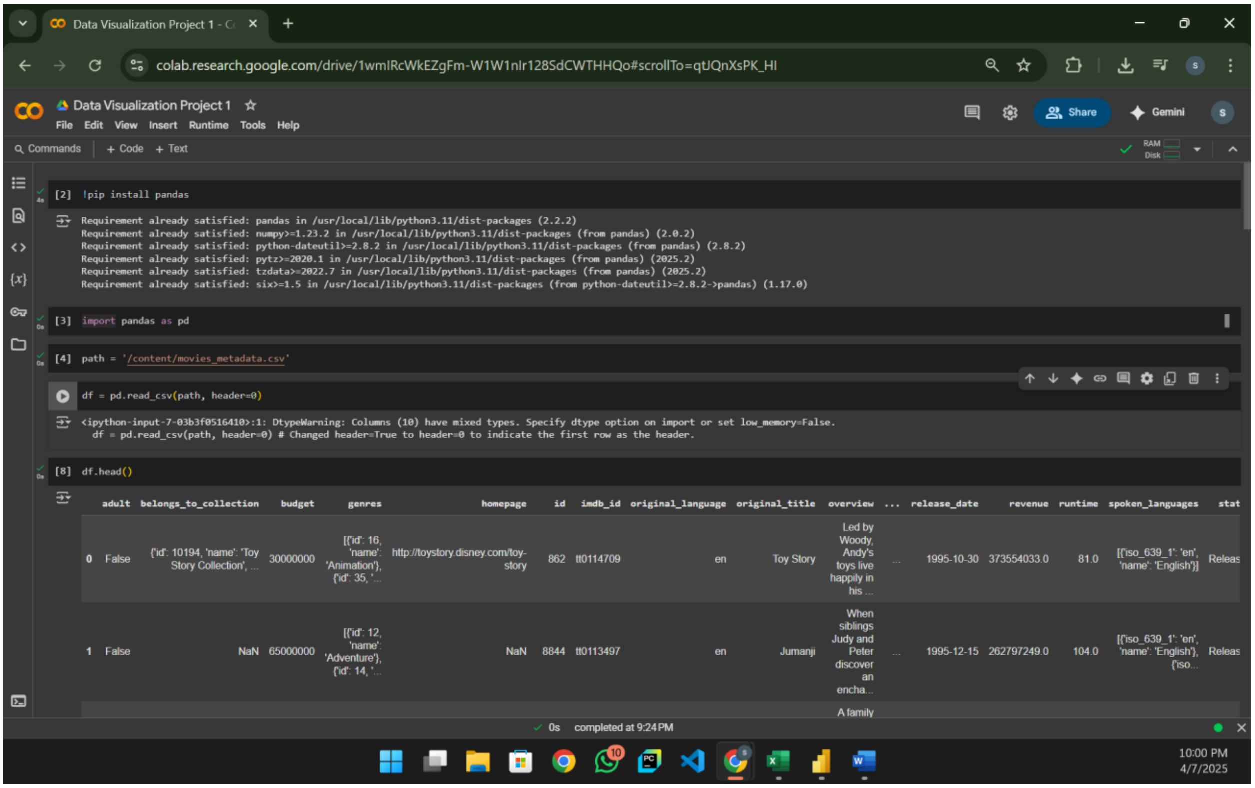
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
adult	belongs_tc	budget	genres	homepage	id	imdb_id	original_la	original_tit	overview	popularity	poster_pat	productior	productior	release_dε	revenue	runtime	spoken_la	status	ta
FALSE	{'id': 10194	30000000	[{'id': 16, 'n	http://toys	862	tt0114709	en	Toy Story	Led by Wo	21.94694	/rhlRbceoE	[{'name': 'F	[{'iso_316ε	########	3.74E+08	81	[{'iso_639_	Released	
FALSE		65000000	[{'id': 12, 'name': 'Adve		8844	tt0113497	en	Jumanji	When sibli	17.01554	/vzmL6fP7	[{'name': 'T	[{'iso_316ε	########	2.63E+08	104	[{'iso_639_	Released	Ro
FALSE	{'id': 11905	0	[{'id': 10749, 'name': 'F		15602	tt0113228	en	Grumpier (	A family we	11.7129	/6ksm1sjK	[{'name': 'V	[{'iso_316ε	########	0	101	[{'iso_639_	Released	Sti
FALSE		16000000	[{'id': 35, 'name': 'Com		31357	tt0114885	en	Waiting to	Cheated o	3.859495	/16XOMpE	[{'name': 'T	[{'iso_316ε	########	81452156	127	[{'iso_639_	Released	Fri
FALSE	{'id': 96871	0	[{'id': 35, 'name': 'Com		11862	tt0113041	en	Father of t	Just when	8.387519	/e64sOI48	[{'name': 'S	[{'iso_316ε	########	76578911	106	[{'iso_639_	Released	Jus
FALSE		60000000	[{'id': 28, 'name': 'Acti		949	tt0113277	en	Heat	Obsessive	17.92493	/zMyfPUel	[{'name': 'F	[{'iso_316ε	########	1.87E+08	170	[{'iso_639_	Released	A L
FALSE		58000000	[{'id': 35, 'name': 'Com		11860	tt0114319	en	Sabrina	An ugly du	6.677277	/jQh15y5Y	[{'name': 'F	[{'iso_316ε	########	0	127	[{'iso_639_	Released	Yo
FALSE		0	[{'id': 28, 'name': 'Acti		45325	tt0112302	en	Tom and H	A mischiev	2.561161	/sGO5Qa5	[{'name': 'V	[{'iso_316ε	########	0	97	[{'iso_639_	Released	Th
FALSE		35000000	[{'id': 28, 'name': 'Acti		9091	tt0114576	en	Sudden De	Internatior	5.23158	/eoWvKD6	[{'name': 'U	[{'iso_316ε	########	64350171	106	[{'iso_639_	Released	Te

ratings.csv (From Kaggle – TMDB dataset):

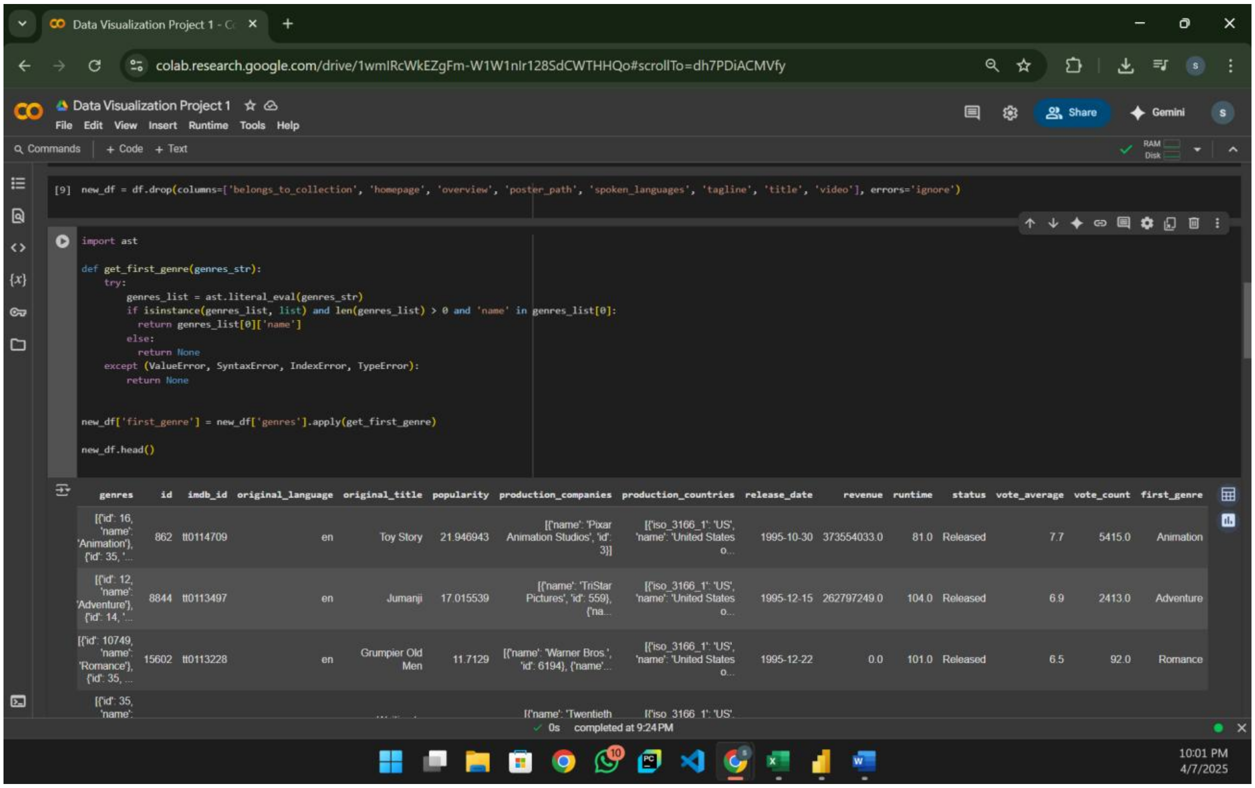
This dataset includes over 100,000 user ratings for movies. Each rating includes a user ID, movie ID, rating score, and timestamp which allows us to evaluate user behavior, preferences, and seasonal trends based on the ratings.

userId	movieId	rating	timestamp
1	110	1	1.43E+09
1	147	4.5	1.43E+09
1	858	5	1.43E+09
1	1221	5	1.43E+09
1	1246	5	1.43E+09
1	1968	4	1.43E+09
1	2762	4.5	1.43E+09
1	2918	5	1.43E+09
1	2959	4	1.43E+09
1	4226	4	1.43E+09
1	4878	5	1.43E+09

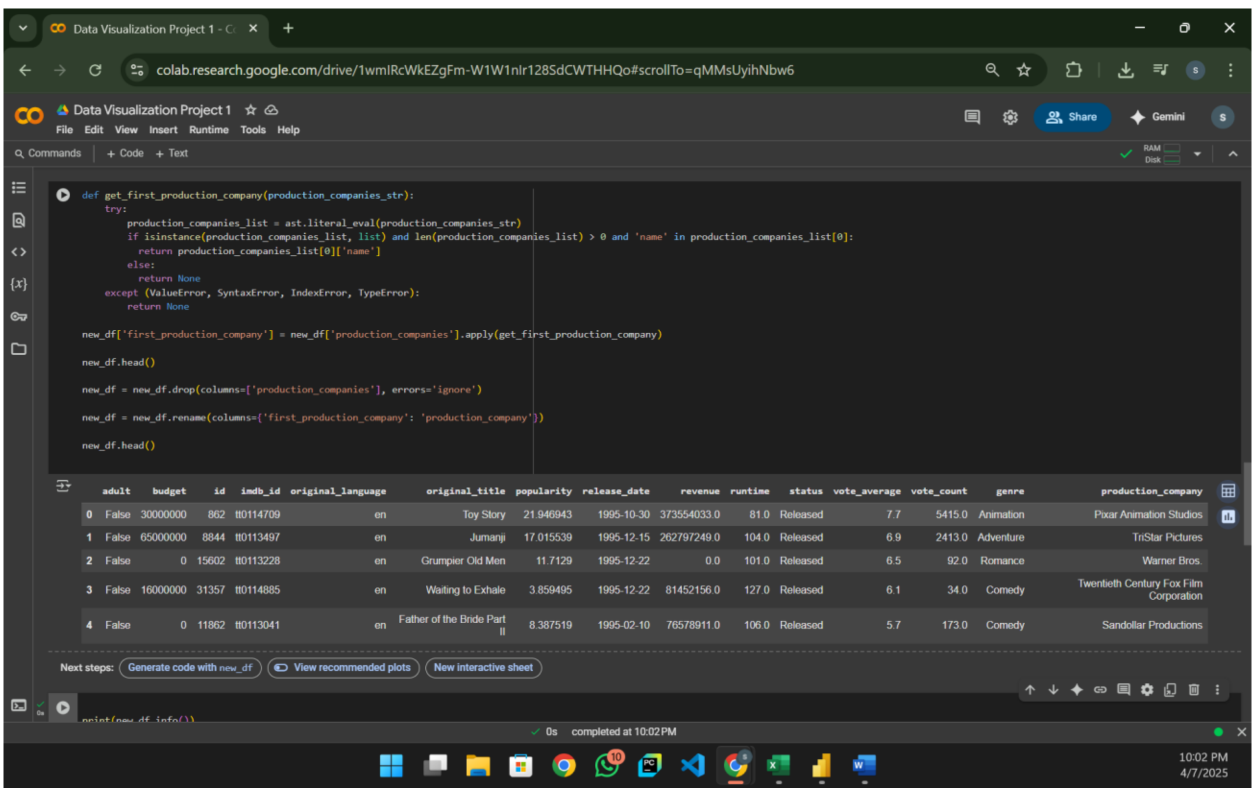
Data Transformation:



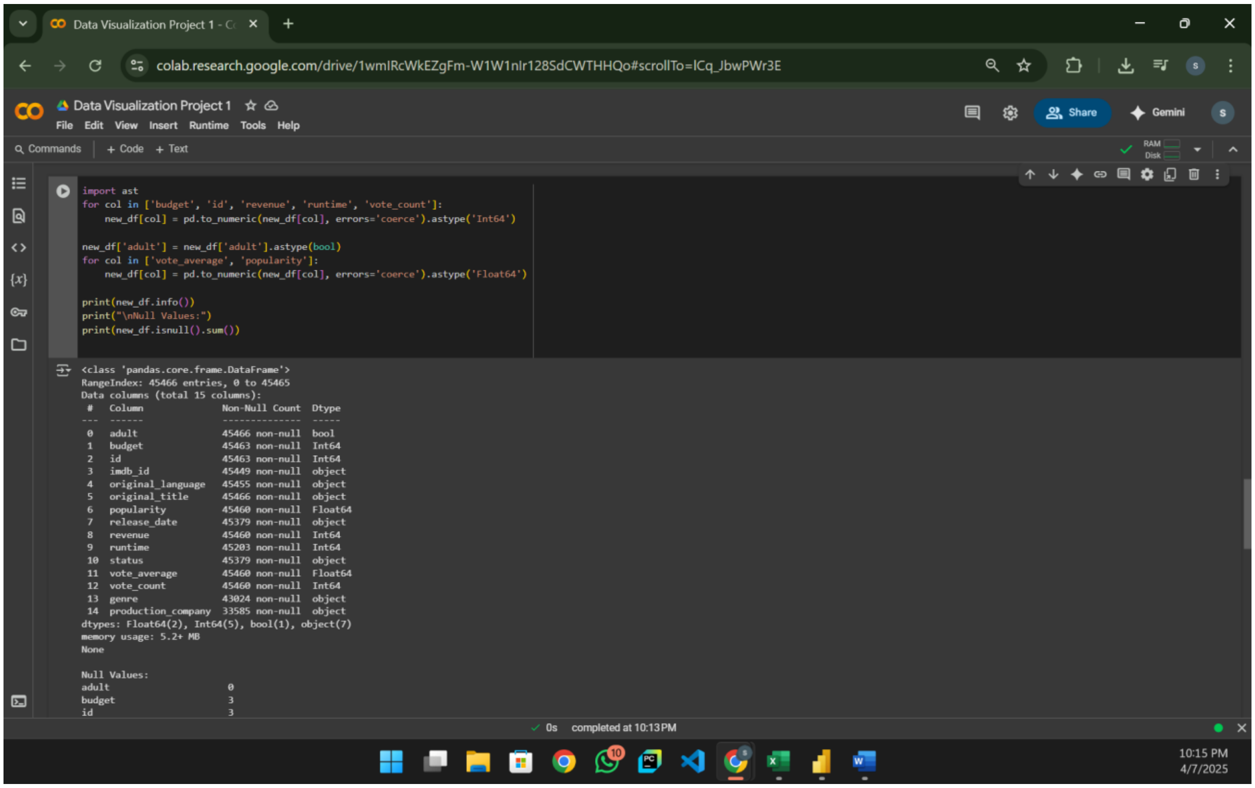
Here we can see that the dataset is loaded using pandas framework and df.head() is called to expose the first five rows. The dataset is not clean and transformed. There are many unwanted columns, null values and nested data.

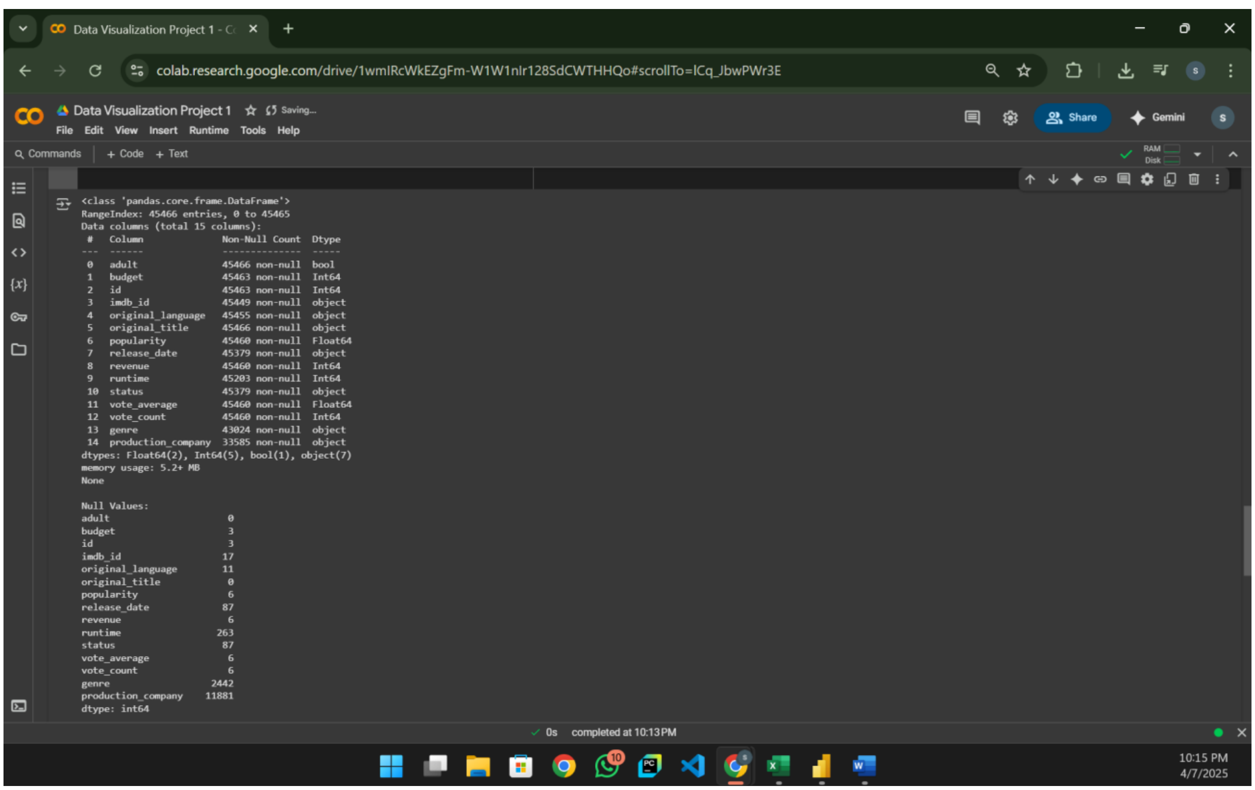


Dropped unnecessary columns and retrieved the genre form the nested list.

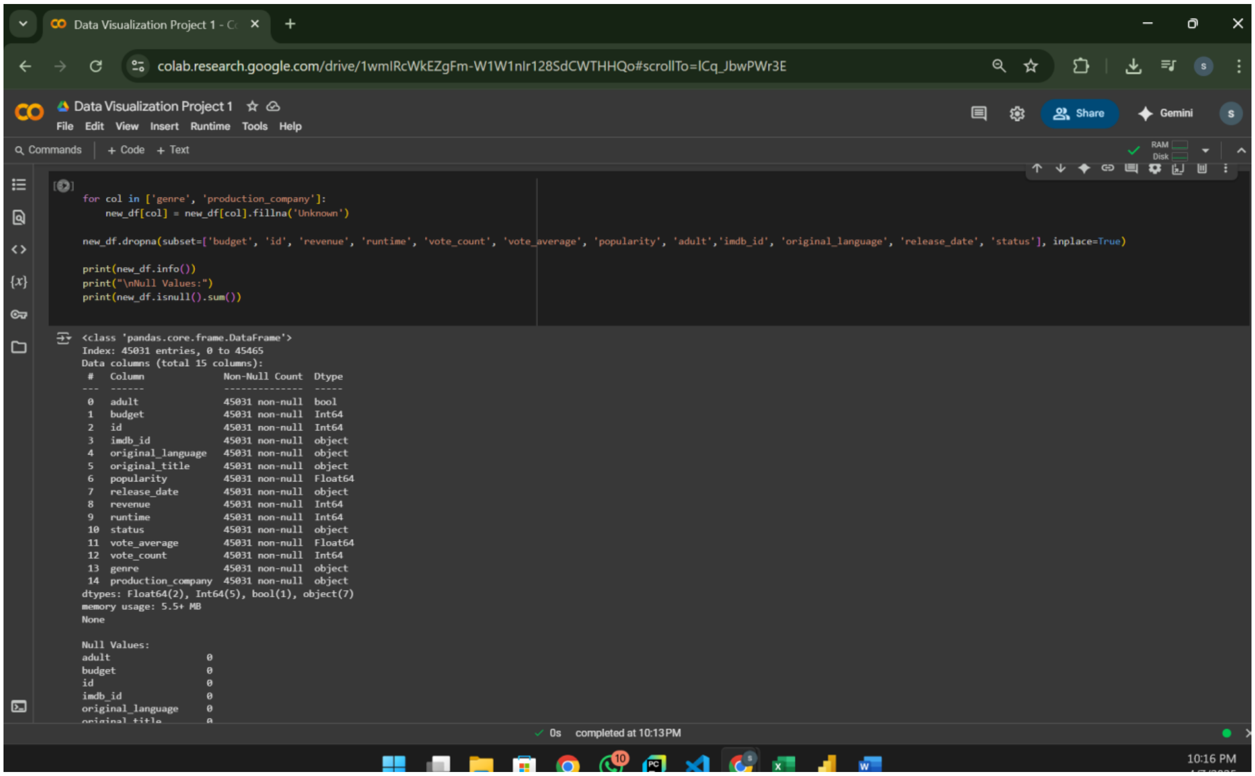


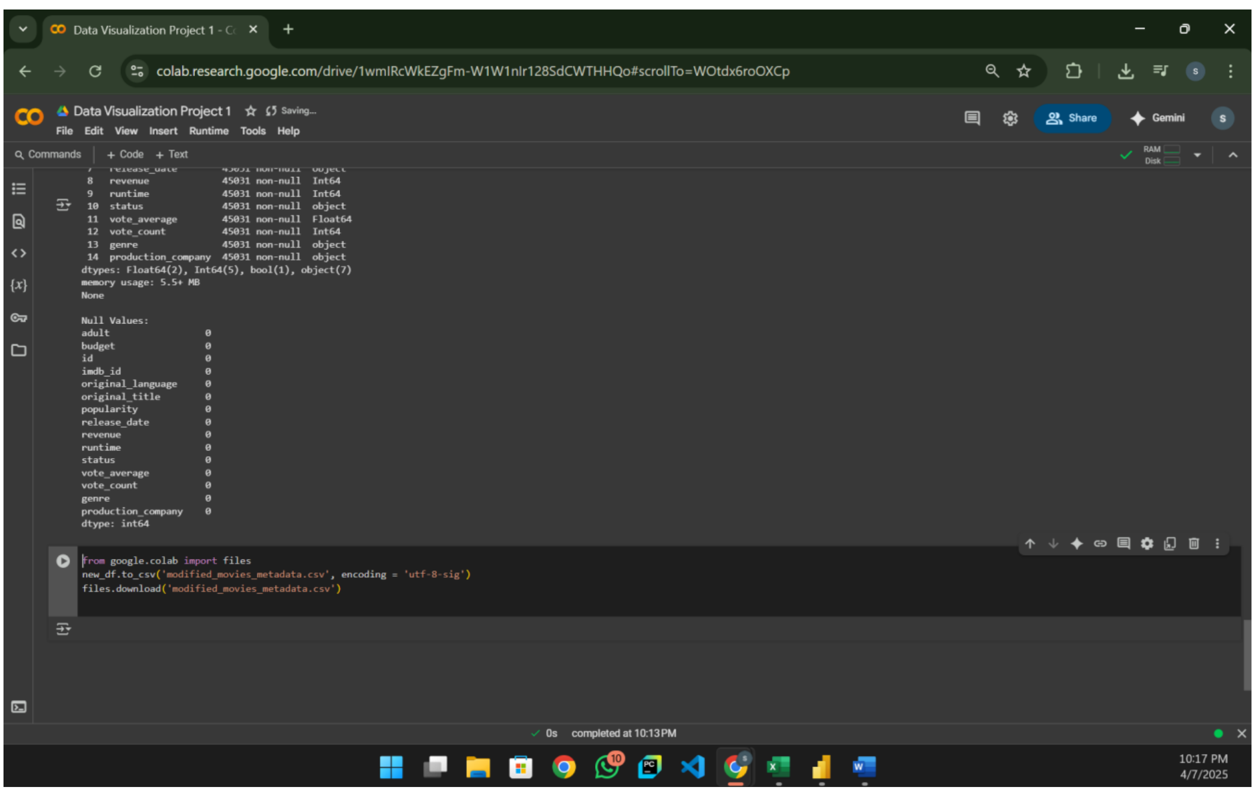
Now we can see that the dataset is cleaned and transformed.





Here we have changed the datatypes of the columns to their appropriate datatypes like int and float. The dataset also have lots of null values, which are handled in the next step.





Here the dataset has no null values because they are dropped. Therefore, our dataset is clean and transformed ready for visualizations.

Target:

We are planning on analyzing:

1. Revenue VS Budget Analysis
2. Top genres by average rating
3. Trending popularity over years
4. Runtime distribution
5. Movies by language
6. Top production companies by revenue
7. Vote count vs vote average

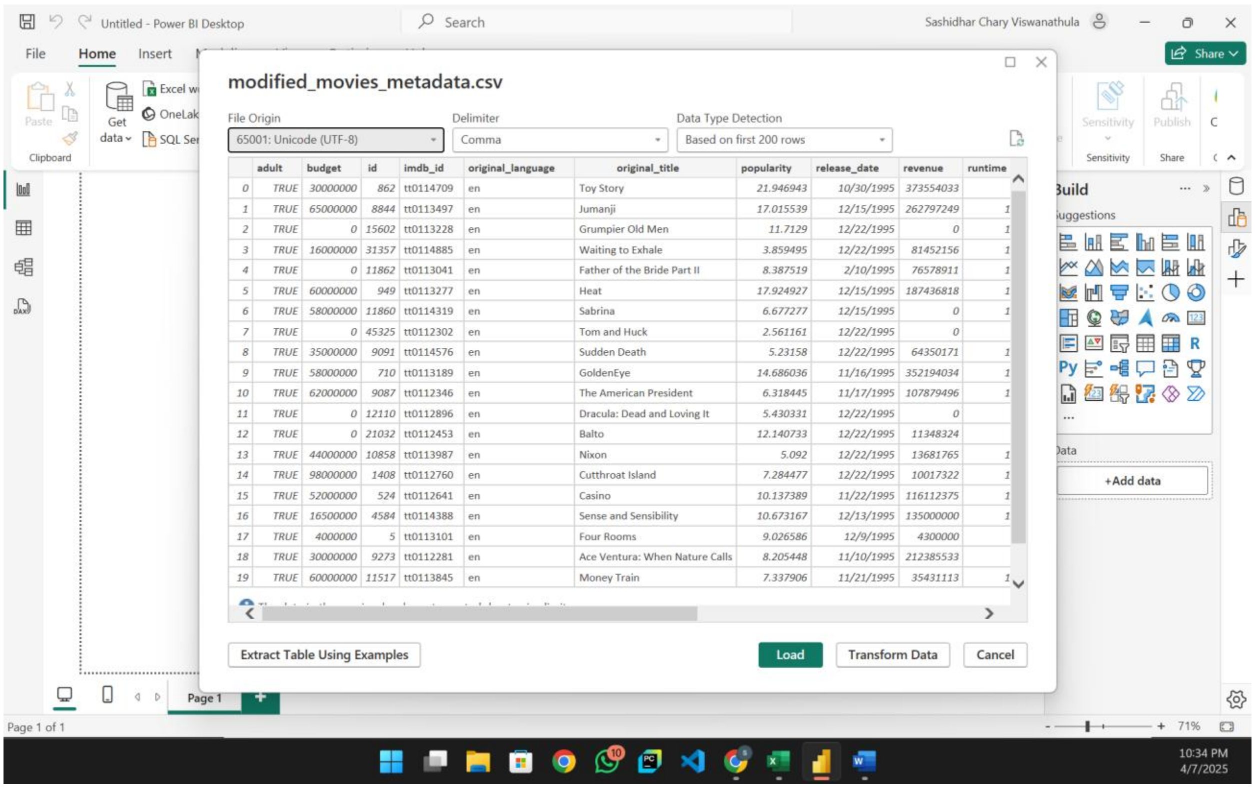
Actions:

1. Filtered the data that doesn't make sense or bring any value to the project. Like null values, unwanted columns etc...
2. Aggregated the columns to get more granularity on the project and much more summarized analysis.
3. Found relation between budget, revenue, release date, popularity.
4. Utilized suitable visualizations to better understand the data. (E.g.: bar chart, scatterplot etc...)

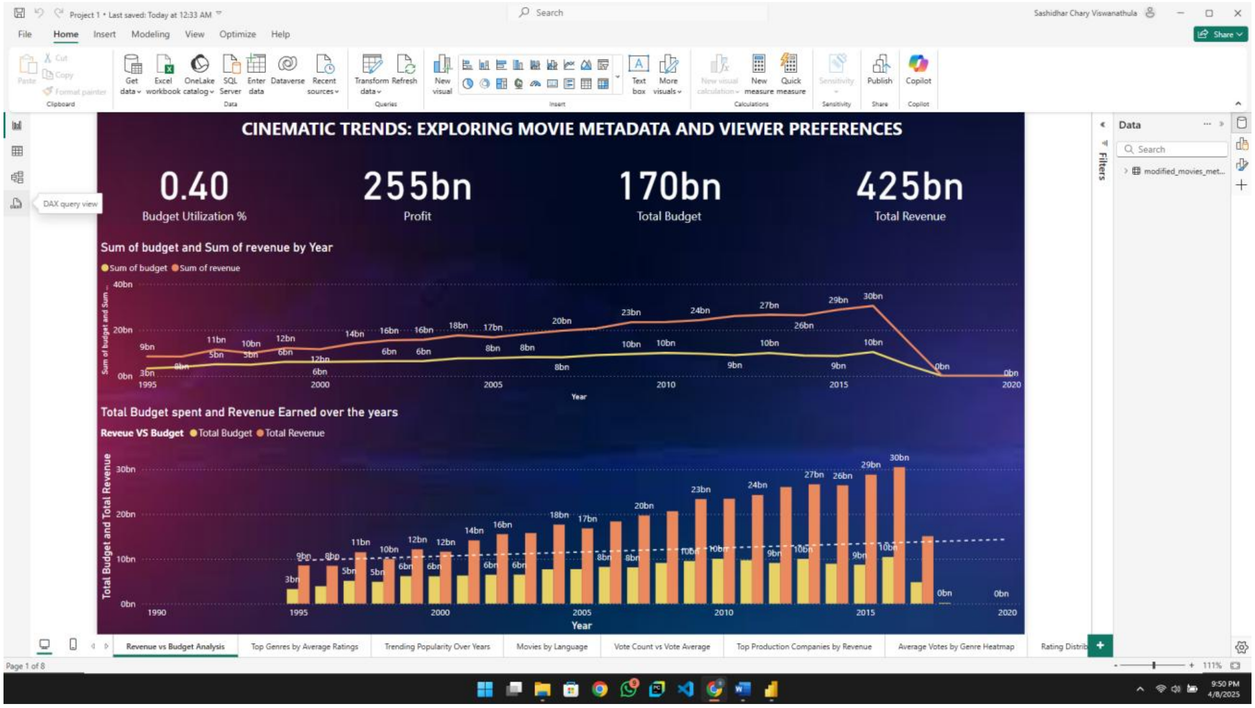
Tool used for visualization:

For this project, we have utilized Power BI to generate visualizations because it is easy to use interface and data loading is easier and flexible in this platform.

Data Loading:



Visualization 1: REVENUE VS BUDGET OVER TIME



This visualization compares total revenue vs total budget per year which is in the bar chart and trends of both overtime in the line chart. This page also has highlights summary KPIs for budget utilization percentage, total revenue, total budget, profit.

KPI's: By seeing the KPIs, we can say that the budget utilization is at 0.4%, which is extremely low, indicating very low budget utilization. The Profit is 255 billion, total budget is 170 billion, total revenue is 425 billion. Despite a low budget utilization percentage, the industry has generated a substantial profit of 255 billion, showing a strong return on investment over the observed period.

Line Chart:

The above chart shows the sum of budget and revenue by year from 1995 to 2020. As we can see, budget (Orange line) remains relatively studied between six billion and 10 billion from 2000 onward and revenue (Pink line) shows an upward trend, peaking sharply at 30 billion in 2018 then dropping to zero in 2020 (Possibly due to Covid 19 shutdowns).

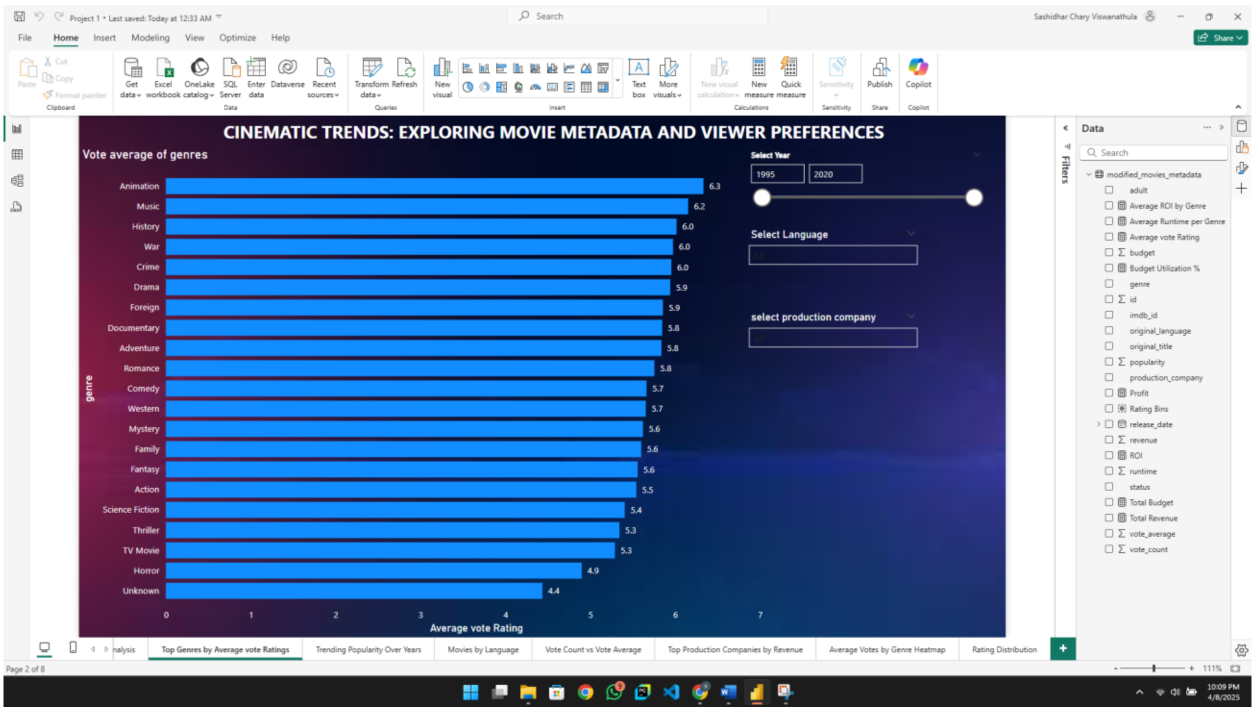
Stacked Column Chart:

The above stacked column chart shows revenue vs budget over the years yellow bars represent budget orange bars represent revenue. There is a visible gap between the two which is the two bars. Revenue bars Are always higher, signaling profit every year. The trend becomes more significant post 2005, showing increased revenue even though budgets grow only slightly.

Insights:

By observing this PBI report we can say that the movie industry has maintained a healthy financial trend over the years, with profits increasing significantly relative to budget. The 30 billion revenues in 2018 was the highest performing year the impact in 2020 is worth highlighting.

Visualization 2: VOTE AVERAGE OF GENRES



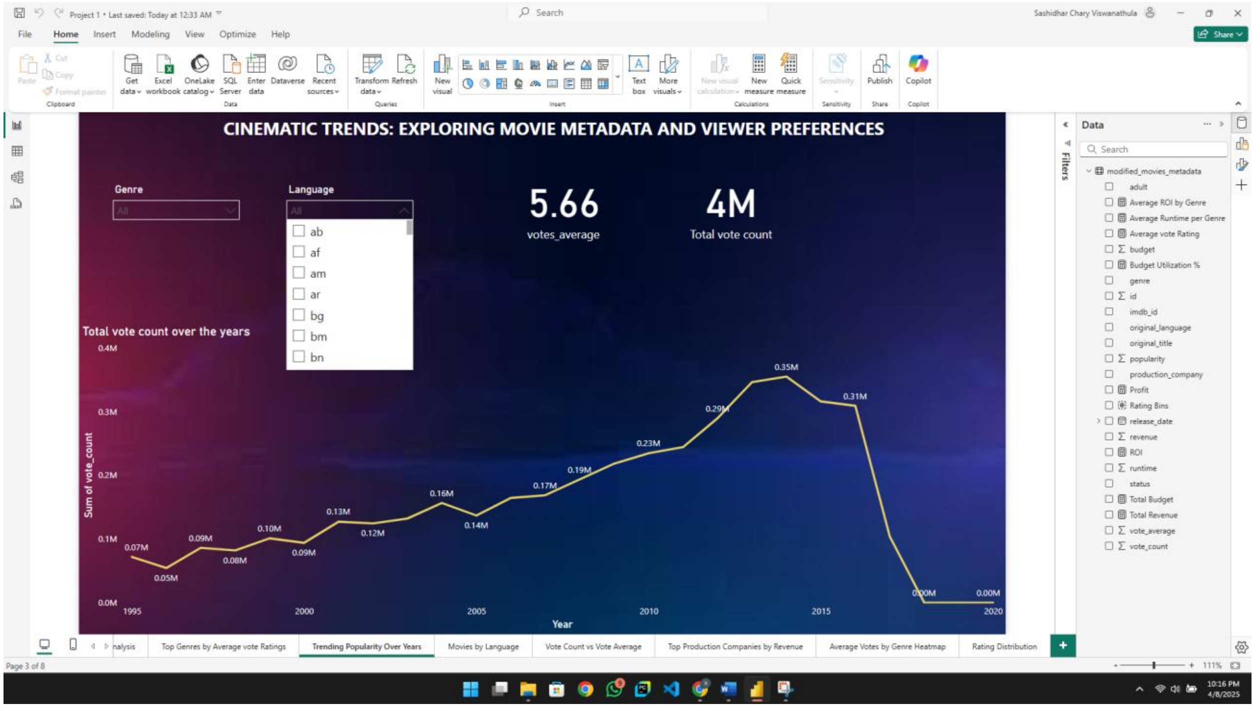
This visualization represents average voting votes of genres. This is a horizontal bar chart displaying the average vote rating for each movie genre we have used a dax measure called **average vote rating = AVERAGE (modified_movies_metadata[vote_average])**. The slicers (Year, language, production company) Remind the same, enabling interactive filtering for deep exploration.

Insights:

On the X axis we have average vote rating and on the Y axis we have genre. This chart displays the average audience rating for each movie genre based on voting data from the data set it allows us to compare how positively or natively weas have rated different genres on average.

By seeing the visualization, we can say that animation has a rating of 6.3, music 6.2, history or war or crime 6.0 are the highest rated genres. TV movie is 5.3 and thriller 5.3 are underperforming, possibly due to lower budgets, production quality or niche appeal.

Visualization 3: TRENDING POPULARITY OVER YEARS



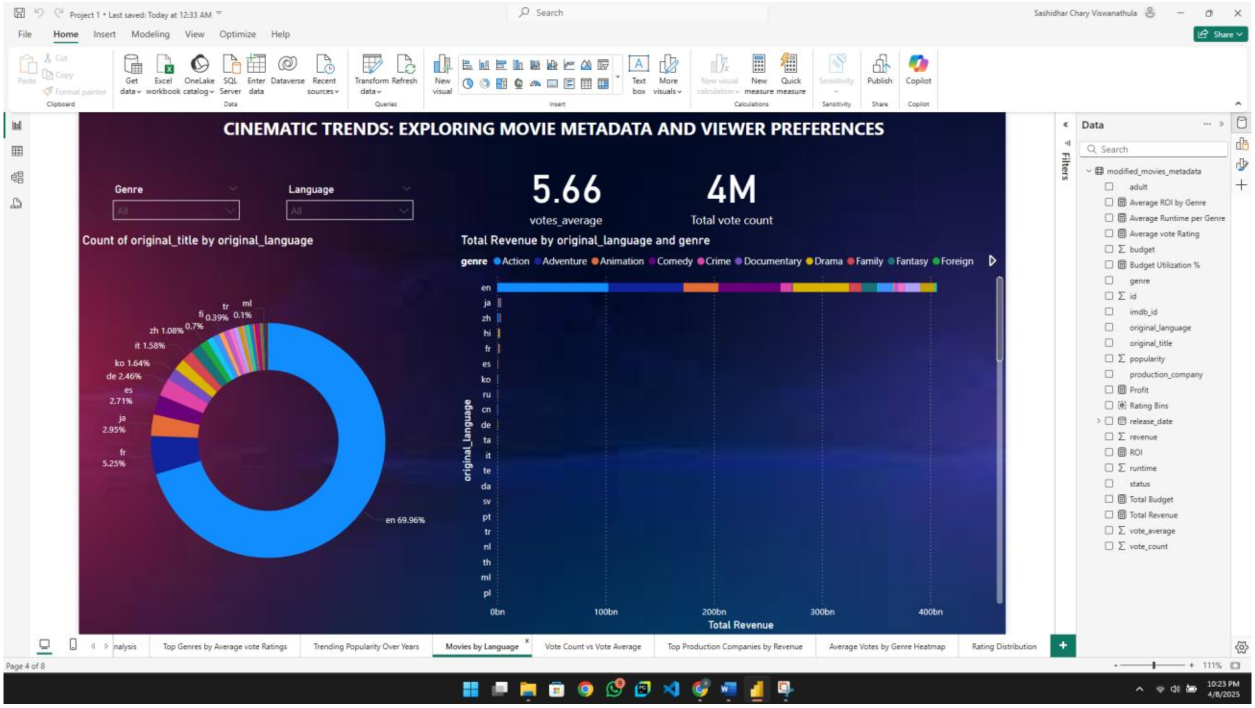
The above line chart shows the sum of vote count by year from 1995 – 2020. Here I've utilized two KPIs which are average board rating and total vote count and two filters which are genre and language.

By observing the chart, from 1995 to 2005 a steady slow growth in both counts from 0.04M to 0 .12M which tells us that the digital movie platforms or metadata sources were limited in audience interaction. From 2006 to 2014 a significant spike from 0.16M to a peak of 0.35M votes in 2014 says that there is a sharp price in increased digital engagement, due to the rise of online movie databases, streaming platforms or mobile access to content from 2015 to 2018 vote counts remain relatively high indicating sustained viewer activity from 2019 to 2020 a dynamic drop to zero votes this is most likely due to disruptions like the COVID 19 pandemic which stalled movie releases and audience interaction.

Insights:

We can say that there is a clear correlation between year and digital audience engagement watch rise steadily with Internet and streaming growth. The sudden drop in 2019 to 2020 is because of the global impact of COVID 19 the inclusion of genre and language filters alert us for detailed investigation into how specific categories or language based films trended in popularity overtime.

Visualization 4: MOVIES BY LANGUAGE



In this Visualization we have used two charts, one is doughnut chart, and one is bar chart. Represents a distribution of movie count by the original language in which they were produced. There are also two KPIs here, one is average vote rating, other is total vote count.

Donut Chart: Percentage of language:

By observing the donor chart, we can say that English is the dominant language accounting for 69 .96% of all movies. Other languages like Hindi are 5.2%, French 2.9%, Japanese 2.7%, Spanish 2.4%, Korean 1.6% and rest other languages are less than 1%.

This shows a heavy lean towards English language films, which is common in global movie datasets however non-English films do show up in notable volumes, especially from Bollywood.

Bar Chart: Revenue by language:

English not only dominates in volume but also in total revenue reaching over 400 billion non-English languages such as Hindi Japanese Spanish contributes smaller but still visible revenue figures. Each genre is color coded to show the difference.

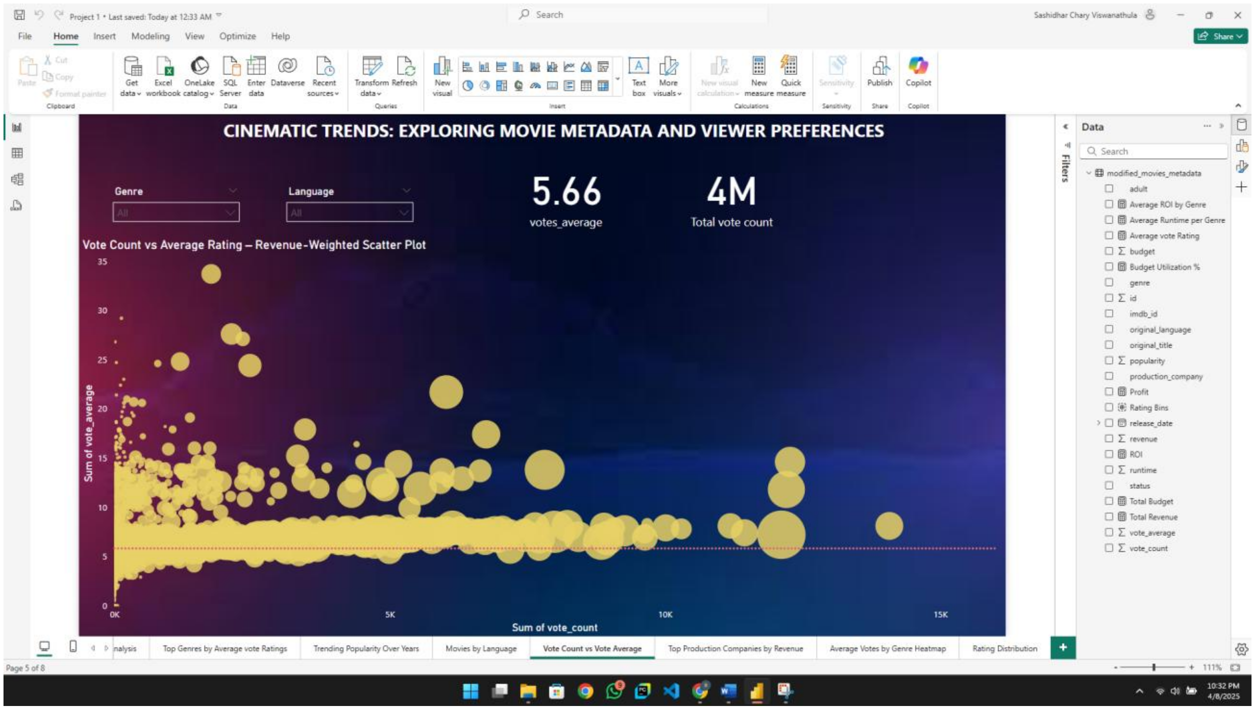
English language films dominate both quantity and profitability, but diverse journals and substantial revenue exists in global markets too, especially in highly specialized genres.

Insights:

This Visualization provides a clear global distribution view for content acquisition teams, showing where the majority of content originates from. Marketing or streaming platforms, helping tailor regional catalogs. production studios, Identifying untapped language - genre revenue pairs.

Visualization 5:

VOTE COUNT VS AVERAGE RATING – REVENUE-WEIGHTED SCATTER PLOT



The above chart is a scatter plot on X axis we have sum of word count on Y axis we have sum of vote average bubble size is proportional to revenue. Disk slatter plot allows us to analyze three performance metrics simultaneously.

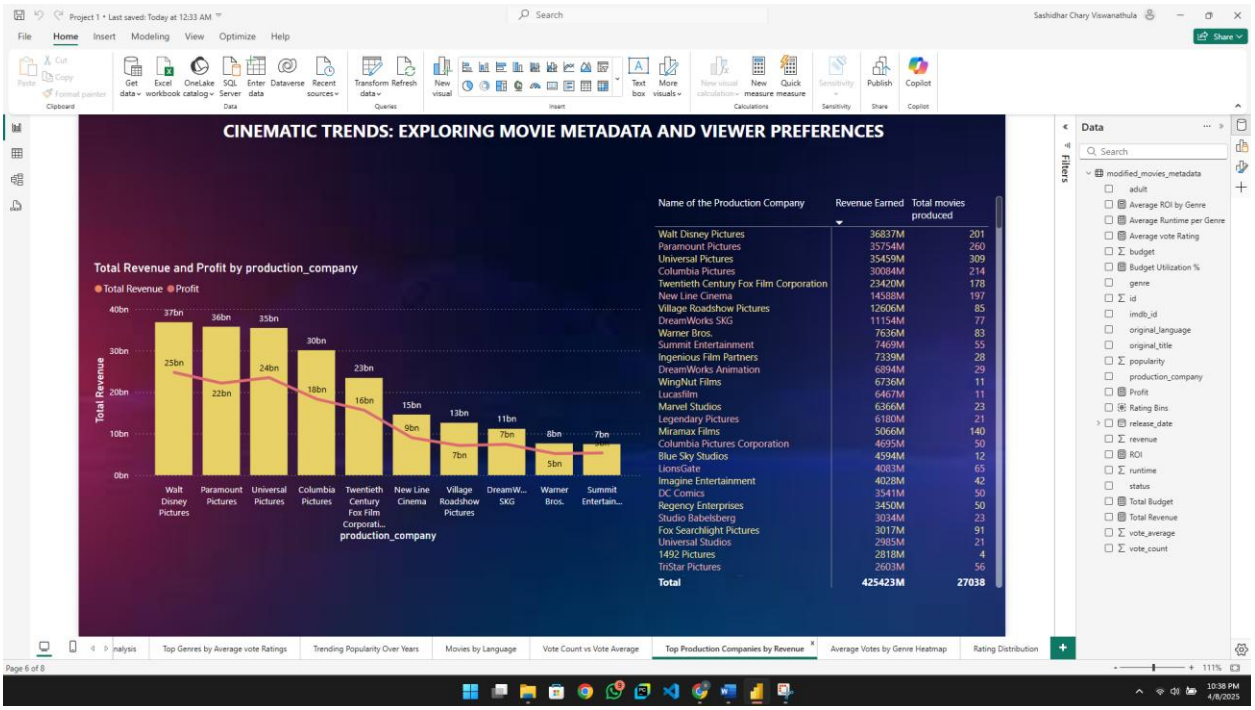
1. How many people voted for a movie?
2. How highly did those people rate the movie?
3. How much revenue does the movie make?

Visually we can observe that a majority of bubbles cluster around the mid lower Y axis indicating a concentration of average rated movies rating between five and 10. The horizontal spread of bubbles shows that some movies have very high vote counts exceeding 15,000 larger bubbles are generally on the right side which tells us the movies with more votes tend to generate modest revenue.

Insights:

By observing this scatterplot, we can say that high revenue is not equal to high ratings because many large bubbles are around the average rating 5 to 7. Success of a film lies in the mid rating but high vote count zone mainstream blockbusters that receive wide attention but average ratings high rating, low vote low revenue films suggest hidden gems or niche productions loved by a small but passionate audience.

Visualization 6: TOP PRODUCTION COMPANIES BY REVENUE



This visualization is a combination of bar and line chart. X axis we have production company on the Y axis we have total revenue, and the line indicates profit this chart compares revenue vs profit for top production companies. The table visualization has revenue earned in millions and total movies produced

The chart tells us Walt Disney Pictures has a revenue of 37 billion with a profit of 25 billion. By this we can say that Disney is the undisputed leader with strong margins and a large catalog of 201 movies. Paramount pictures and Universal pictures are two production companies whose revenue is close to Disney and profit is slight lower than Disney around 22 billion to 24 billion. which accounts for 260 plus movies from each production company. As a chart moves to the right side revenue and profit both decline but profit decreases more steeply than revenue for example Summit Entertainment and Warner Bros make 7 billion and 8 billion, with profits of one to 2 billion, suggesting higher costs or low ROI.

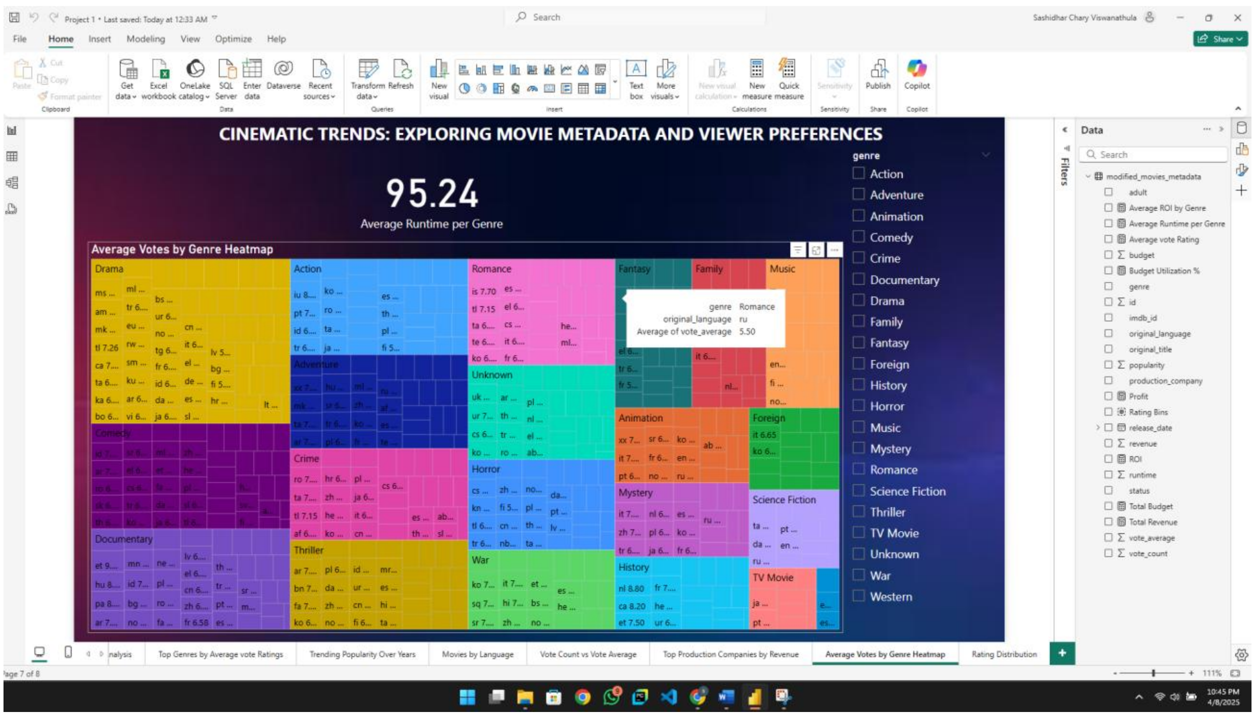
Matrix visualization adds important context by showing volume of production such as universal pictures leads in total movies of 39 dream works animation with fewer titles of 29 studios like

Marvel Studios and legendary pictures appear mid table with fewer movies but relatively high revenues, showing quality over quantity success.

Insights:

By observing the whole visualization, we can say that Walt Disney's dominance in both volume and profitability confirms its position as a market leader although there is a noticeable correlation between scale and success but not always fuse studios with small catalogs deliver strong returns companies with higher output but lower profits might need operational or content strategy adjustments.

Visualization 7: GENRE-LANGUAGE HEATMAP



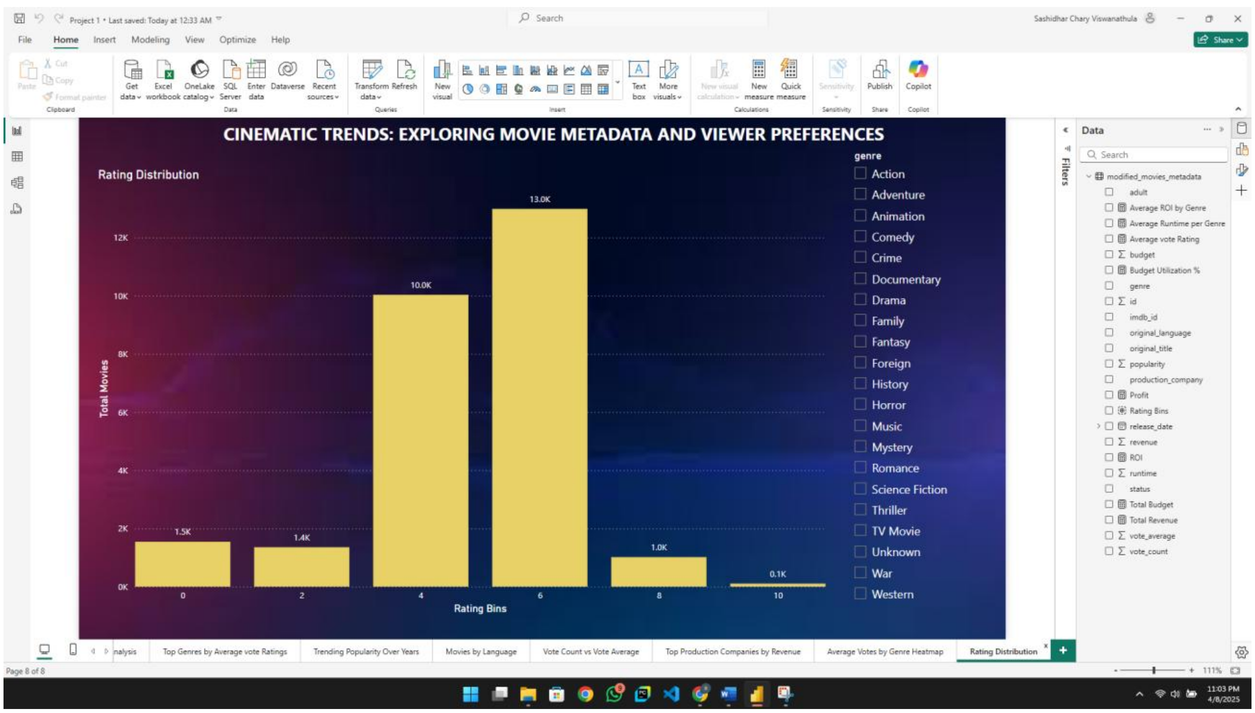
The above visualization is a heat map which shows average votes by genre. In category axis we have genre and original language, and the metric displayed here is voting average. This heat map is a multilayered analysis which shows how each genre performs in terms of audience rating and which original language versions of the genre received higher or lower ratings. Each small tile represents a unique combination of genre plus language and is proportional in size based on the number of titles and displays the average via rating for that combination.

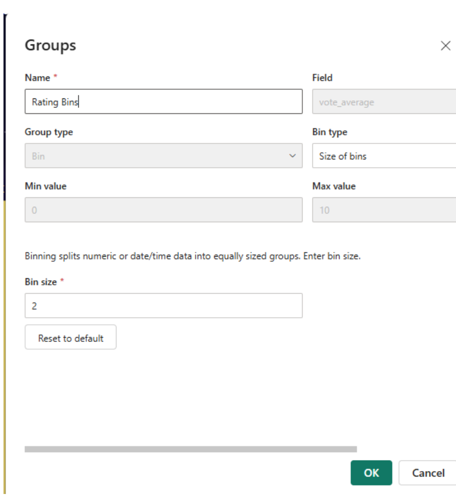
Insights:

By observing the heat map, we can say that drama has the widest representation, indicating high volume and variety across languages like animation, foreign and documentary also show healthy ratings across multiple languages. For example, in the tool tip romance movies in Russian average a 5.5 rating which shows moderate performance foreign genre entries often hover above 6.5, showing strong audience appreciation for international cinema.

Some genres like TV movie horror tend to have lower ratings or fewer language variations. Crime thriller and mystery genres also display many language tiles with ratings close to the global average of 5.66.

Visualization 8: RATING DISTRIBUTION





This visualization shows the rating distribution on the X axis we have ranges like 02468 and 10 on the Y axis we have the number of movies that fall into each rating bin. For this visualization I've binned the vote average field with the size of two, ranging from 0 to 10. The rating buckets are 0-2,2-4,4 -6, 6 - 8, 8 - 10 and exactly 10.

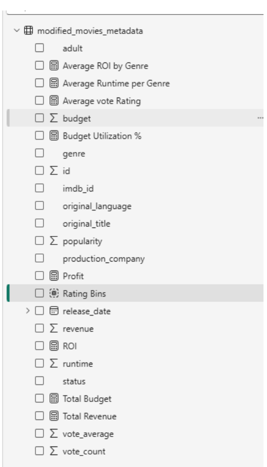
Insights:

By observing the chart, most movies are rated between 4 to 8, which captures 23,000 of the total movies out of 27,000. Specifically, 6 to 8 range is the peak, with 13,000 movies rated here. Is also strong with 1000 entries. Only 1000 movies scored between 8 to 10 and even fewer 100 hit a percent score of 10. Similarly, only 1500 scored between zero to two indicating few movies are universally disliked. Ratings show a bell-shaped distribution towards the center.

Storytelling Approach:

Stage 1: understanding and Exploring the Dataset:

What does raw data look like?



The dataset, as seen in the Fields pane, is from a table called modified_movies_metadata. It contains attributes such as:

Identifiers: id, original_title, imdb_id

Descriptive: genre, original_language, production_company

Performance metrics: budget, revenue, vote_average, vote_count, runtime, ROI, Profit, Budget Utilization %

Calculated fields: Average ROI by Genre, Rating Bins, etc.

What key variables and trends stand out?

Revenue & Profit Trends: Movies have consistently generated strong profits, with a total revenue of 425bn and profit of 255bn. Walt Disney, Paramount, and Universal are the top earners.

Viewer Sentiment: Average rating is 5.66 across 27,038 movies. Majority of movies fall in the 6-8 rating bin, indicating a large volume of moderately liked content. Animation, Music, and History genres received the highest ratings.

Language & Genre: English (69.9%) dominates, but foreign-language films also show strong niche performance. Performance of genres varies by language. e.g., Romance in Russian vs. Comedy in Korean.

Viewer Engagement: Vote count peaked around 2014-2018 and dropped off in 2020 (likely due to pandemic). High vote count doesn't always equal high ratings, popularity and quality are not strictly correlated.

Stage 2: Data Processing and Correlation Analysis

In this stage, we move from descriptive analysis to uncovering underlying patterns and relationships between variables. Key operations include:

- Cleaning missing or inconsistent values
- Merging rating and metadata tables
- Creating new fields (e.g., Rating Bins, ROI, Profit)
- Grouping and binning numeric fields for trend analysis

Correlation Analysis:

Vote Count vs Rating vs Revenue (Bubble Chart): High revenue movies often lie in the mid-rating zone (5–7), not necessarily the highest-rated. Some well-rated movies have low vote counts, revealing undiscovered gems.

Genre vs Average Rating: Top genres in ratings: Animation, Music, History**.** Bottom genres: TV Movie, Horror. Supports genre-based content curation or investment strategies.

Runtime vs Rating: Longer movies may have better storytelling and thus higher ratings. To be validated in later phases with scatter or trend lines.

Production Companies: Revenue vs Profit: Walt Disney leads with high revenue and profit efficiency. Marvel Studios shows high ROI with fewer productions. Smaller companies tend to have lower revenue-per-movie.

Work Management:

As mentioned in the proposal below, tasks have been completed for the project 1:

- Explored metadata features like genres, average ratings, popularity
- Detected basic correlations such as genre vs. average rating, release year vs. popularity
- Identified outliers and missing data
- Generated initial visual dashboards

Data Cleaning & Transformation:

- Handle missing or inconsistent values in the movies_metadata.csv file (e.g., missing runtimes, differently formatted release dates).
- Convert data types (e.g., release dates to datetime, budget/revenue to numeric).

Exploratory Data Analysis (EDA):

- Analyze the distribution of movie genres, release years, and popularity scores.
- Visualize the average user rating across different genres and decades.
- Create basic visualizations using Power BI

Initial Correlation Analysis:

- Identify relationships between variables such as:
- Movie runtime vs. average rating
- Budget/revenue vs. popularity
- Number of ratings vs. movie rating score

Name of the person	Responsibility taken	Percentage distribution
Sashidhar Chary Viswanathula	Worked on data cleaning, transformation, two visualizations and documentation	40%
Nithish Karanam	Worked on four visualizations, IEEE paper, PPT and documentation	30%
Vidya Reddy	Worked on two visualizations and documentation	30%