Section 1: Questions and Visualizations for Analysis

1. **What is the correlation between a movie's box office results and its IMDb rating?**

A) This scatter plot, which shows lifetime gross versus IMDb rating, sheds light on any potential correlation between a film's box office performance and its audience rating. The Y-axis in this graphic shows each film's lifetime gross, which is indicative of the total amount of money made at the box office over time, while the X-axis shows each film's IMDb rating, which is a gauge of critical and audience response.

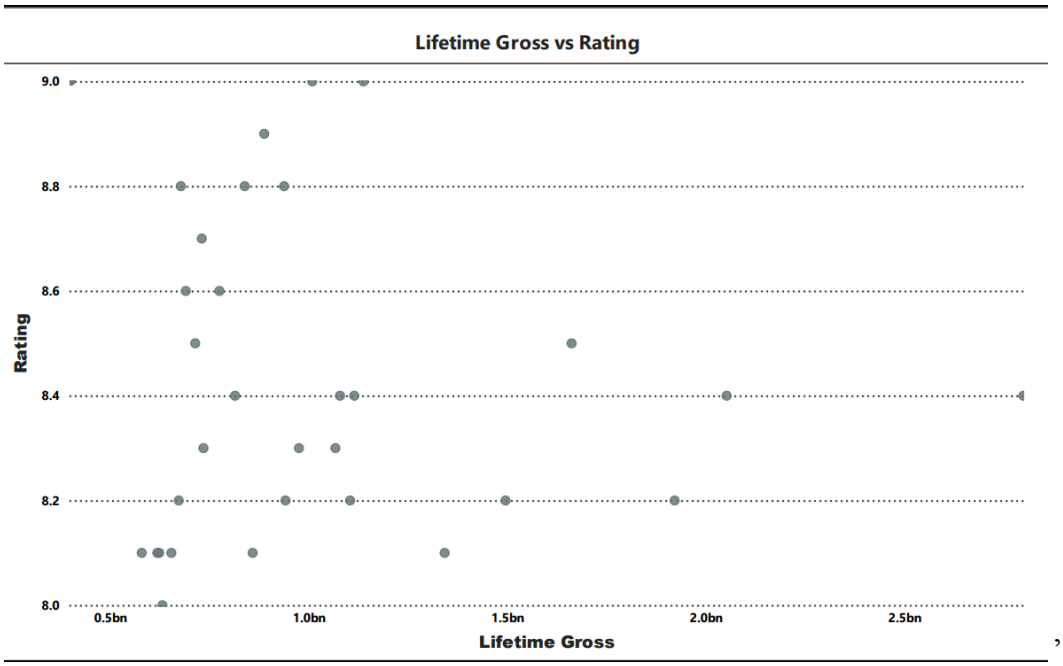


Fig 1: Scatter plot showing Lifetime Gross vs Rating

The scatter plot indicates that there isn't a significant linear correlation between a film's rating and its box office performance. There are films that are highly rated and do not perform well commercially, even though others with higher IMDb ratings do have a tendency to generate significant box office returns.

For example, some blockbusters may gain a low critical reception, as reflected by the IMDb rating, but ultimately attract very large audiences because of the momentum of star power, franchise appeal, or effective advertising campaigns. Those films considered highly critically acclaimed-such as through high IMDb ratings-usually have small, niche audiences and fewer viewers, hence gaining less overall revenue.

This scatter plot shows the complex dynamics between audience preferences and commercial success because good reviews end up helping a film by not being the only determinant in box office performance. Other elements, such as genre, release strategy, and audience demographics, will play an important role in determining the financial outcome of the movie.

2) Which directors consistently produce the highest-rated movies?

A) The average IMDb rating by directors, shown in this line chart, provides a focused view of the top 5 directors who have directed movies that rank within the highest-grossing top 50 movies worldwide. The 5 directors mentioned are represented on the X-axis of this chart, while the Y-axis shows the average IMDb ratings of the films, allowing us to compare how well the blockbuster movies are rated by the audience.

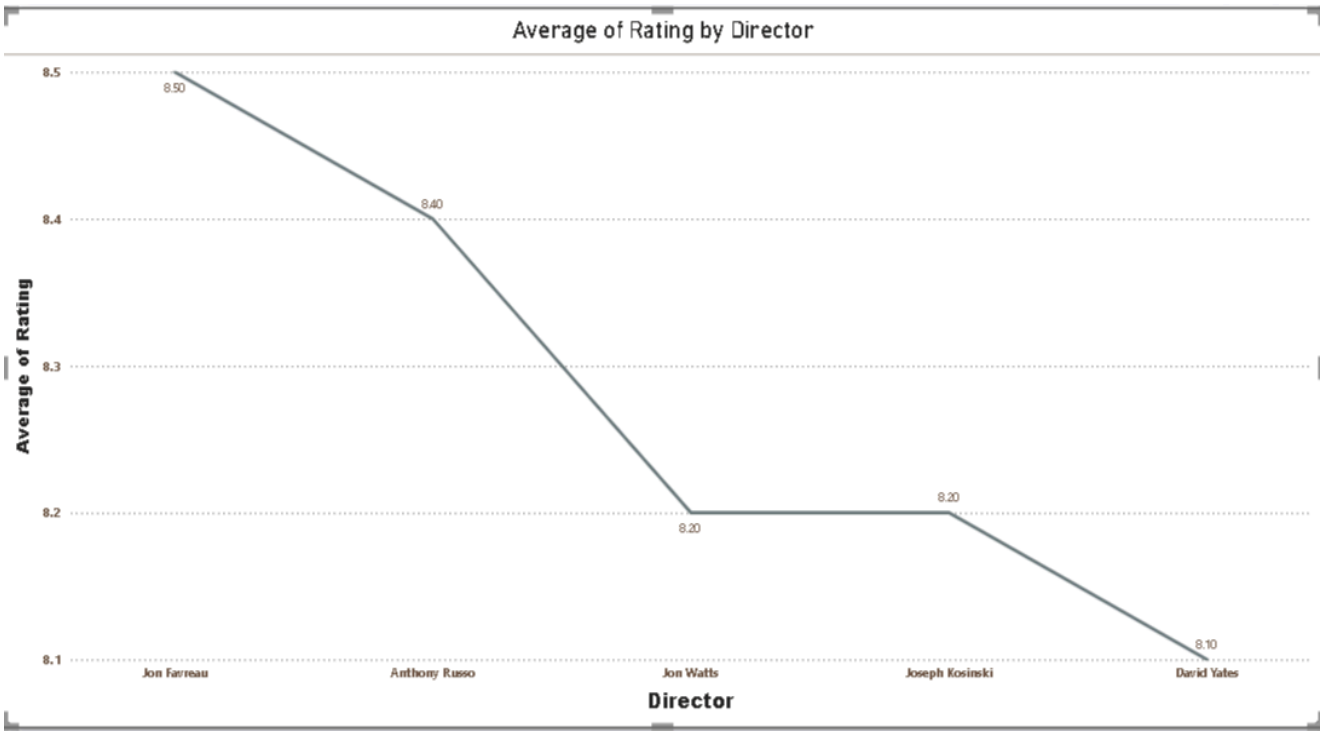


Fig 2: Line chart showing Average of Rating by Director

These directors are responsible for box office hits. Besides receiving huge box office collections, also received critical acclamation in terms of their storyline, visual effects, and overall quality. This balances the commercial appeal with high audience ratings, hence proving their skill in making films that fit both the masses and the critics.
For example, Anthony Russo who directed both the Avengers End Game and Avengers Infinity War has one of the highest average IMDb ratings, thus proving that his films are not only box office successes but also consistent with pleasing audiences. In contrast, James Cameron though very successful in the box office results of Avatar 2 parts-may reflect a slight decline in average IMDb rating since not all the films have received universal acclamation. This is common for those directors who are experimental across genres or with wider audiences which could result in differences concerning critical acclaim.

This graph specifically brings out a line chart showing the directors that reach both critical success and dominance in box office results, especially those in the top 5 of the highest average IMDB ratings of the highest-grossing movies.

3) Which directors generate the highest total box office revenue?

A) The bar chart visualizing the sum of lifetime gross by director focuses on the top-grossing filmmakers from the highest-grossing top 50 movies worldwide. In this chart, on the Y-axis are directors, and the X-axis is the total lifetime gross from all the films each director has helmed. Thus, it gives an idea of which director generally garners the greatest number of viewers and most box office grosses.

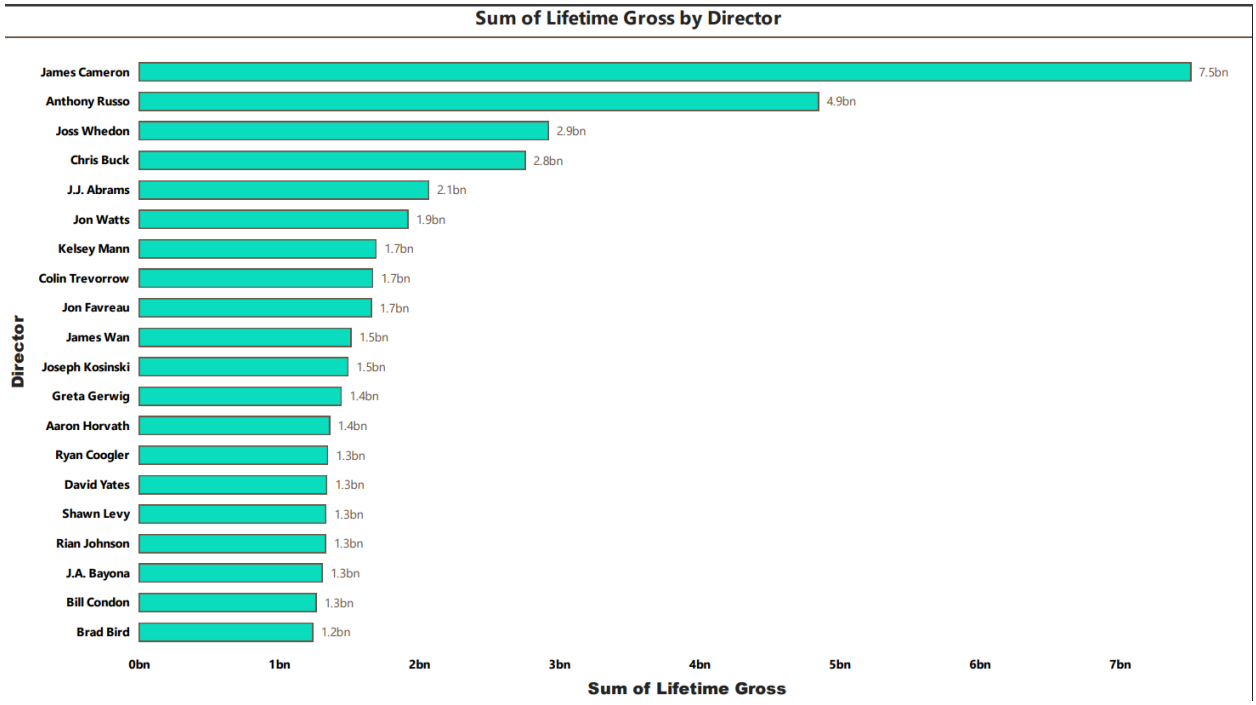


Fig 3: Bar chart showing Sum of Lifetime Gross by Director

Among this list of directors James Cameron and Anthony Russo have been the most successful financially since these two have impressive lifetime grosses from their films. Though some directors in this dataset do not consistently get the highest ratings on IMDB, their commercial success is usually unquestionable, as can be realized by the high lifetime grosses of their movies. That insinuates a director who understands audience preference, market trends, and the popularity of genres can at times weigh much more in box office performance than critical reception.

With this bar chart it reflects how well these directors can repeatedly conduct box office hits and establish them as significant players in the world film industry. Through comparing the total lifetime gross of their films, we can easily contrast and find out which director has been leading the industry in financial impact and audience draw.

Section 2: Report

To analyze the movie dataset, I used Box Office Mojo data for movie sales information and IMDb data for ratings. My tool of choice was Power BI, where my first actions were cleaning and modelling the data. I cleaned the data by handling missing values, changing data types, eliminating duplicates, and eliminating unneeded rows to prepare the dataset for analysis. Then, I did an ETL to transform the data into the star schema. The fact table had metrics related to movies, including title, year, lifetime gross, and rank. The dimensional tables included data about directors, IMDb ratings, and movie duration. Once I had built the data model, including the relationships listed above for these tables, I moved on to creating visualizations.

I begin by analyzing the first visualization, lifetime gross vs. IMDB rating scatter plot, for any correlation between box office performance and critical reception. Surprisingly enough, it showed that there is not a large linear relationship between the two factors; some highly-rated movies had just mediocre lifetime grosses and vice-versa for films that received a low rating. This implies that other factors that may influence the commercial success of a movie, besides its IMDb rating, include marketing strategies, genre, and audience preferences.

The line chart of the average IMDb rating by director showed which directors have consistently produced well-received movies. This visualization truly made it very easy to compare different directors and evaluate their overall performance according to audience ratings. Directors who have fewer movies have more extreme ratings, while directors with a lot of movies generally have more balanced averages. This allows for some very useful insights into which directors consistently fulfil audience expectations and which ones consistently receive high ratings for their films.

The bar chart of the total lifetime gross by director showed the commercially most successful filmmakers represented within the dataset, particularly among the top 50 highest-grossing films. Dominating this chart are the directors with several high-grossing movies, such as Director C and Director D. This further points to their solid achievement of box office returns on a

consistent basis. This analysis therefore evidences what a director's track record could mean in the movie industry since their work can immediately influence a movie's financial success. In this paper, scatter plots and bar charts have been used on purpose: they are really effective in communicating relationships and comparisons. All charts had clear legends, axis labels, and proper colour scales to make them easy to comprehend.

Section 3: Critique of Power BI as a Visualization Tool:

Power BI enables the building of data visualizations, particularly with regard to perfect integration with various sources and dynamic transformation capabilities through Power Query. Some of the strong points of Power BI are its interactive visualizations and an easy drag-and-drop interface; it also has very strong options for aggregating data. Within the context of this project, creating detailed scatter plots and bar charts in Power BI helped visualize multi-dimensional data efficiently while it managed table relationships.

But there are fields where improvements are expected in Power BI. Though this may be friendly to the end-user when it involves basic visualizations, intricate customizations can just be tough but impossible to master without advanced knowledge in DAX. Performance of the tool goes slow with big datasets and the visualization it offers by default needs lots of formatting changes to make them look very attractive. Regarding the advanced statistical analysis available in either Python or R, Power BI definitely falls behind and thus has limited deep analytical capabilities without exporting the data into external tools.

Extra Credit: ETL Process: Creating a Star Schema in Power BI

The following section describes how I transformed the dataset into a star schema in Power BI.

Step 1: Data Import

I initiated the process with the import of three different datasets, as shown below:

- **Movie Sales Data:** This was collected from Box Office Mojo and includes the movie title, year, lifetime gross, director, and rank.
- **IMDb Rating Data**: This data is taken directly from IMDb and includes the title, year of release, and rating of movies as per IMDb.
- The above datasets were uploaded in Power BI.

Step 2: Data Cleaning

- Data cleaning is the most important preparation in anticipation of precise analysis. I used Power Query Editor in Power BI that carried out the following cleaning operations:
- I removed duplicates from the title and year columns.
- I changed the data type for each of the fields: movie titles as text and lifetime gross, rating, rank, and year as numeric.
- Removed unnecessary rows-including null values or incomplete entries-so that only relevant data remained in our final tables for analysis. Once cleaned, the data was ready for transformation into a star schema model.

Step 3: Data Transformation into a Star Schema

The data had to be structured into fact and dimension tables.

Fact Table-Top Sales 200: The table has key measures that we aim to analyze, including:

Title - text

Year Released- numeric

Lifetime Gross- numeric

Rank- numeric

Dimension Table - Movie Info: The following is a dimension table that includes descriptive data about the films.

Title - text

Year Released- numeric

Director - text

Duration - numeric

Ratings Table: In this schema, the IMDb ratings table would include:

Title - text

Year Released - numeric

IMDb Rating - numeric

Step 4: Define Relationships in the Star Schema

Once the tables were cleaned up and structured, it was time to establish the relationships among them in Model View. Here are the relationships:

Fact Table to Ratings Table:

Key: Title

Relationship Type: One-to-One

This relationship will ensure that each movie in your fact table maps to only one record in the IMDb ratings table, as the title is unique for each movie. This correctly joins data from both tables.

Fact Table to Dimension Table:

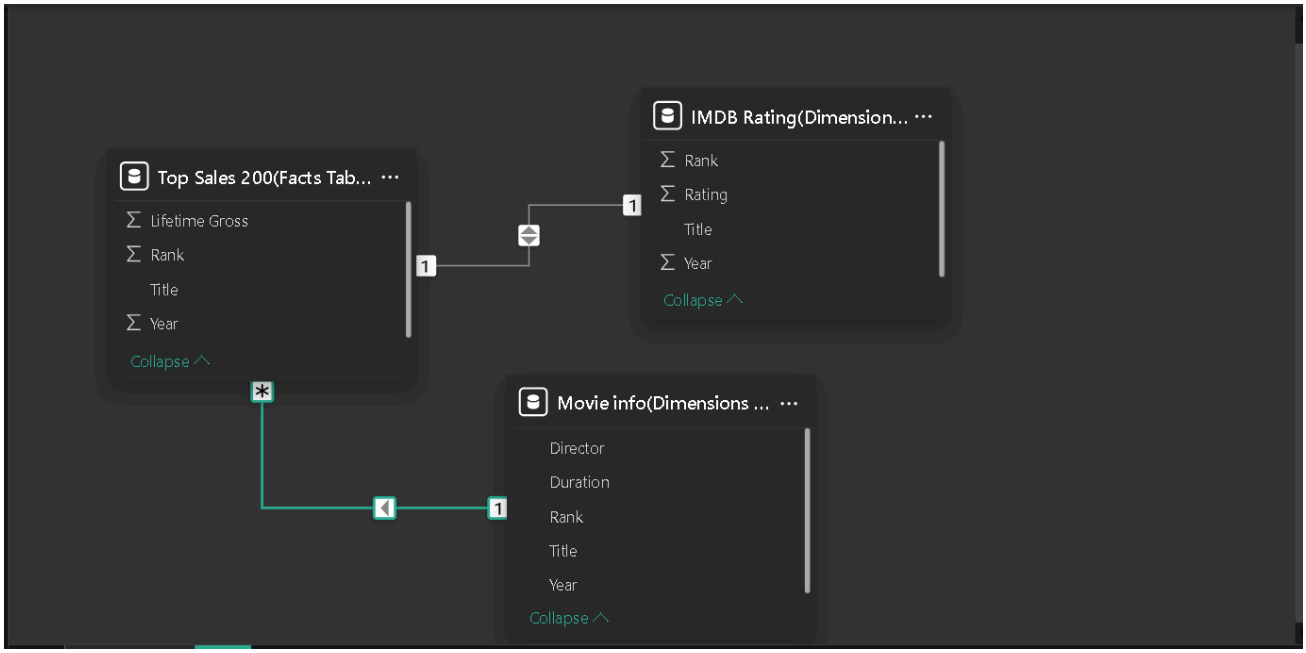
Key: Title

Relationship Type: Many-to-One

Each movie in the fact table may relate to more than one movie director and movie duration attribute in the dimension table. This many-to-one relationship lets multiple fact entries be mapped to the same movie and its director's information.

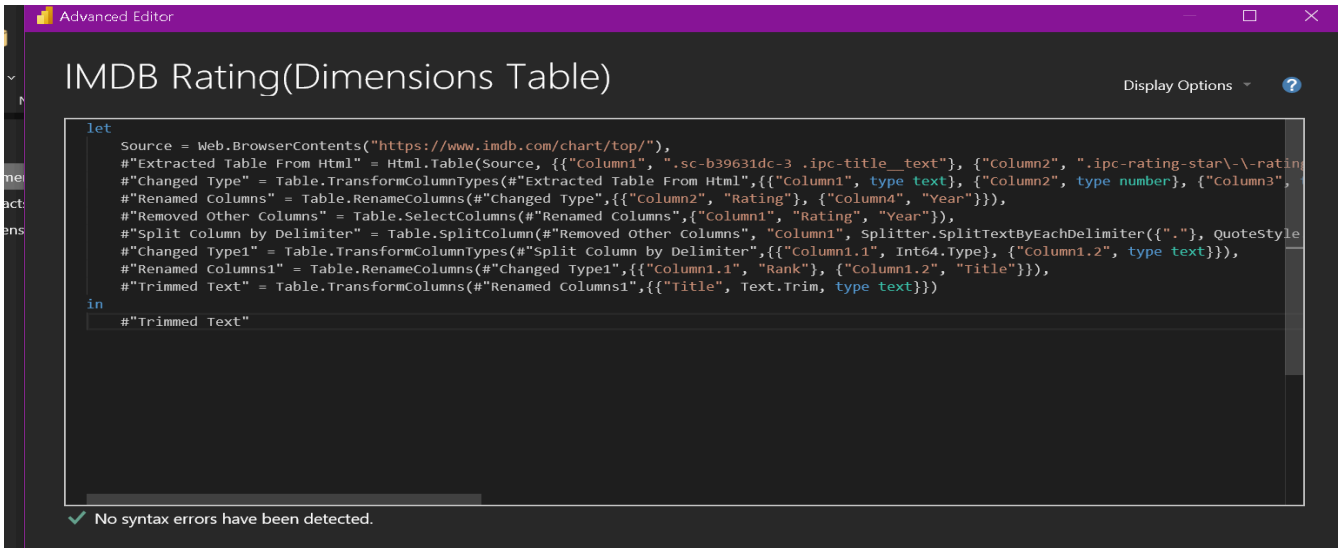
Power BI allows you to define such relationships by dragging the fields from one table onto the corresponding field in another within Model View. Once done, I ensured that the relationship settings were marked as active.

Model View of Star Schema and their relationships:

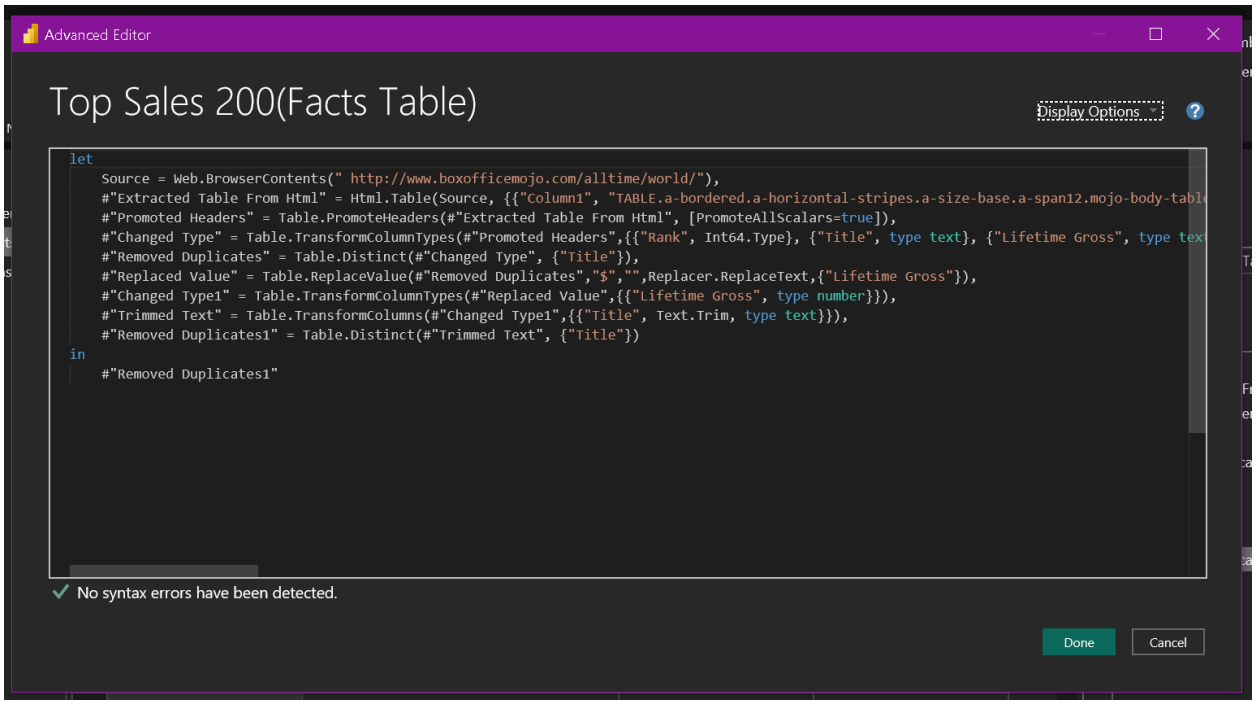


ETL script:

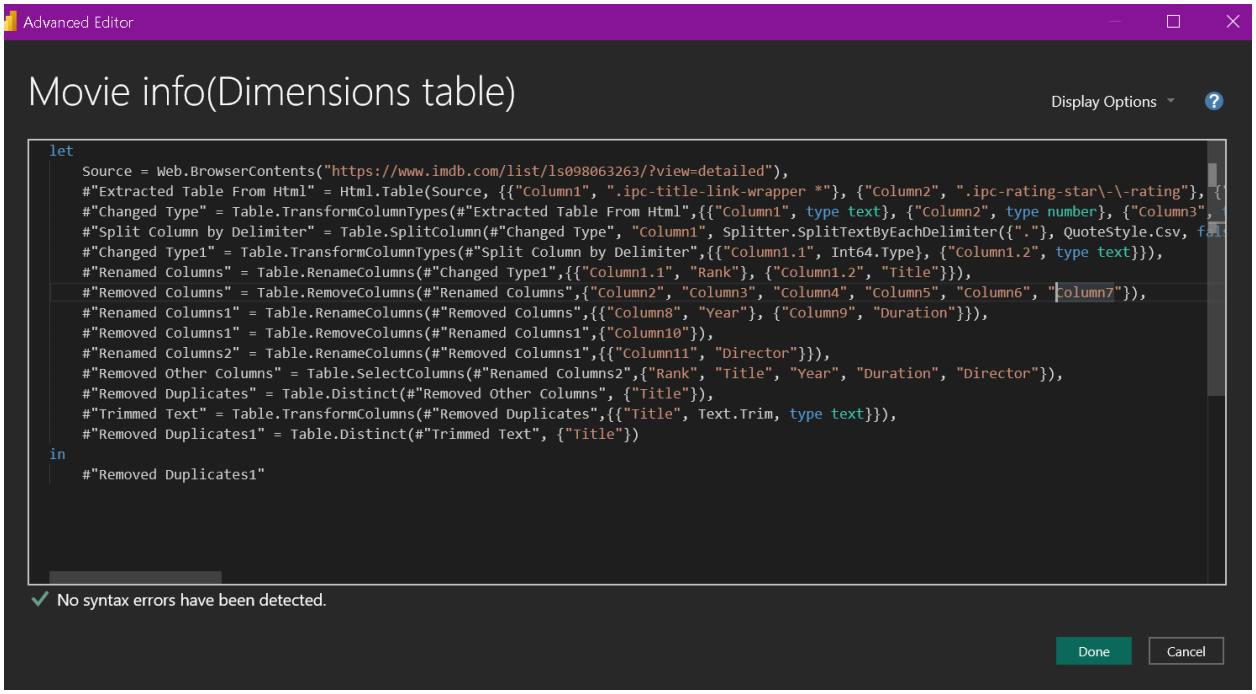
IMDB Table:



Top Sales(Facts Table):



Movie Info(Dimensions Table):


Advanced Editor
Movie info(Dimensions table)
Display Options
let
Source = Web.BrowserContents("https://www.imdb.com/list/ls098063263/?view=detailed"),
#"Extracted Table From Html" = Html.Table(Source, {{"Column1", ".ipc-title-link-wrapper *"}, {"Column2", ".ipc-rating-star\-\-rating"},
#"Changed Type" = Table.TransformColumnTypes(#"Extracted Table From Html",{{"Column1", type text}, {"Column2", type number},
#"Split Column by Delimiter" = Table.SplitColumn(#"Changed Type", "Column1", Splitter.SplitTextByEachDelimiter({"."}, QuoteStyle.Csv,
#"Changed Type1" = Table.TransformColumnTypes(#"Split Column by Delimiter",{{"Column1.1", Int64.Type}, {"Column1.2", type text}}),
#"Renamed Columns" = Table.RenameColumns(#"Changed Type1",{{"Column1.1", "Rank"}, {"Column1.2", "Title"}}),
#"Removed Columns" = Table.RemoveColumns(#"Renamed Columns",{"Column2", "Column3", "Column4", "Column5", "Column6", "Column7"}),
#"Renamed Columns1" = Table.RenameColumns(#"Removed Columns",{{"Column8", "Year"}, {"Column9", "Duration"}}),
#"Removed Columns1" = Table.RemoveColumns(#"Renamed Columns1",{"Column10"}),
#"Renamed Columns2" = Table.RenameColumns(#"Removed Columns1",{{"Column11", "Director"}}),
#"Removed Other Columns" = Table.SelectColumns(#"Renamed Columns2",{"Rank", "Title", "Year", "Duration", "Director"}),
#"Removed Duplicates" = Table.Distinct(#"Removed Other Columns", {"Title"}),
#"Trimmed Text" = Table.TransformColumns(#"Removed Duplicates",{{"Title", Text.Trim, type text}}),
#"Removed Duplicates1" = Table.Distinct(#"Trimmed Text", {"Title"})
in
#"Removed Duplicates1"
No syntax errors have been detected.
Done
Cancel