**1. Data Transformation (40%)**

**1-Clean and structure the dataset (handle missing/null values, duplicates).**  
  
**AMAZON.csv**  
.replaced missing values in director,cast,country columns with unknown.&.remove duplicate rows.

* **Apple.csv**

**.**removed\_duplicates,Replace NULL Values with Average/Median.,In the Transform tab, select Statistics → Average,Transform → Replace Values.

**2-Convert Date into proper Date format and extract Year/Month for analysis.**

* **Amazon.csv**

.changed the type to date.

Duplicated the date column 2 times.

In duplicated column select transform,month/year and rename.(2 columns added).

**3-Categorize movies based on Score (e.g., High: 7+, Medium: 4-7, Low: <4).**

**Amazon.csv**

For this I added a conditional column with these contions-  
3️⃣ **Enter the Conditions One by One:**

| * **Column Name** | * **Operator** | * **Value** | * **Output** |
| --- | --- | --- | --- |
| rating | equals | 18+ | High (7+) |
| rating | equals | R | High (7+) |
| rating | equals | 13+ | Medium (4-7) |
| rating | equals | ALL | Low (<4) |
| rating | equals | TV-Y | Low (<4) |
| rating | is null |  | Unknown |
| rating | does not equal any of the above |  | Unknown |

**Apple.csv**

"IMDB Rating Category".

Set conditions:

If tmdb\_score ≥ 7, then "High (7+)".

If tmdb\_score ≥ 4 and <7, then "Medium (4-7)".

If tmdb\_score < 4, then "Low (<4)".

Else, "Unknown" (for NULL values).

"IMDB Popularity Category".

Set conditions:

If imdb\_votes ≥ 50000, then "High (50K+ votes)".

If imdb\_votes ≥ 10000 and <50000, then "Medium (10K-50K)".

If imdb\_votes < 10000, then "Low (<10K)".

Else, "Unknown" (for NULL values).

**Final\_bollywood.csv:**  
if [score] >= 50 then "High" else if [score] < 20 then “medium” else “low”.

**Final.hollywood.csv:**

if [score] >= 70 then "High" else if [score] < 20 then “medium” else “low”.

**Hotstar.csv:**

Age Rating Age Category

U, U/A 7+, PG → Child

U/A 13+ → Teen

U/A 16+, A → Adult

Unknown → Unknown

1. **Create new calculated fields (e.g., Profit = Revenue - Budget, ROI %).**

**Apple.csv**

Decade = Number.RoundDown([release\_year] / 10, 0) \* 10

Century = Number.RoundDown([release\_year] / 100, 0) + 1

**Final.bollywood.csv,Final.hollywood.csv**

Profit = [Revenue] - [Budget]

ROI% = ([Profit] / [Budget]) \* 100

**Hotstar.csv:**

Running\_Time\_HHMM =Number.ToText(Number.RoundDown([running\_time] / 60)) & "h " &

Number.ToText(Number.Mod([running\_time], 60)) & "m"