**Scenario 1:**

**In the Oscars dataset, some movies appear multiple times due to re-releases. We are programmatically removing duplicate movie entries in the Oscar dataset, where some movies have been re-released (i.e., same movie title appears multiple times with different release years).**

**The logic is dynamic and will auto-handle future changes or additions to the data.**

Some movies like **A Star is Born** may appear multiple times in the Oscar dataset because they were released in:

* Original year (e.g., 1937)
* Re-release (e.g., 1976 and 2018)

You're expected to:

* Identify such duplicates (by movie title)
* Keep **only the latest release year** for each movie – i.e. 2018
* Do this using **Power BI’s Power Query (M)**

let

Source = Csv.Document(File.Contents("C:\Users\Tania Pal\OneDrive\Power BI TMDB Dataset\oscar.csv"),[Delimiter=",", Columns=4, Encoding=65001, QuoteStyle=QuoteStyle.None]),

#"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),

#"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"Film", type text}, {"Year", Int64.Type}, {"Award", Int64.Type}, {"Nomination", Int64.Type}}),

#"Filtered Rows" = Table.SelectRows(#"Changed Type", each [Film] <> null and [Film] <> ""),

#"Sorted Rows" = Table.Sort(#"Filtered Rows",{{"Film", Order.Ascending}, {"Year", Order.Descending}}),

#"Removed Duplicates" = Table.Distinct(#"Sorted Rows", {"Film"})

in

#"Removed Duplicates"

**Data types** in the tmdb\_5000\_movies.csv file from the Kaggle dataset:

| **Column** | **Data Type** |
| --- | --- |
| budget, id, revenue, vote\_count | **int64** |
| popularity, runtime, vote\_average | **float64** |
| genres, keywords, production\_companies, production\_countries, spoken\_languages, cast, crew | **object** (JSON strings) |
| homepage, original\_language, original\_title, overview, poster\_path, release\_date, status, tagline, title, imdb\_id | **text** |

To include movies that are **present in both the TMDB and Oscars datasets**.

I am merging TMDB + Oscar data → deduplicating → parsing nested JSON columns → creating a structured dataset for reporting.

**Steps in Power BI Power Query:**

**1. Load your dataset**

* Go to **Home > Transform Data** to open Power Query Editor.

**2. Identify the JSON column**

* Find the column that holds **JSON strings**, e.g., genres, production\_companies, etc.

**3. Parse the JSON string**

* Go to **Add Column** tab → **Custom Column**
* Use this formula:

Json.Document([genres])

Name this new column something like ParsedGenres

**4. Expand the list (if it's a list of records)**

* If the new column is a **list**:
  + Click the **expand icon** next to the column name (🔽)
  + Select **"Expand to New Rows"**

**5. Expand the records**

* Now the column has individual **records** (objects) like {id=28, name="Action"}
* Click the **expand icon** again → choose fields like id, name
* This gives you columns like genre id, genre name

**6. (Optional) Rename Columns**

* Rename the expanded fields appropriately (genre id, genre name, etc.)

let

Source = Table.NestedJoin(tmdb\_5000\_movies, {"original\_title"}, oscar, {"Film"}, "oscar", JoinKind.Inner),

#"Expanded oscar" = Table.ExpandTableColumn(Source, "oscar", {"Film", "Year", "Award", "Nomination"}, {"oscar.Film", "oscar.Year", "oscar.Award", "oscar.Nomination"}),

#"Replaced Value" = Table.ReplaceValue(#"Expanded oscar",

each [oscar.Year],

each if [oscar.Year] = 28 then 1928

else if [oscar.Year] = 29 then 1929

else if [oscar.Year] = 33 then 1933

else [oscar.Year],

Replacer.ReplaceValue,{"oscar.Year"}),

#"Sorted Rows" = Table.Buffer(Table.Sort(#"Replaced Value",{{"release\_date", Order.Descending}})),

#"Removed Duplicates" = Table.Distinct(#"Sorted Rows", {"original\_title"}),

#"Changed Type" = Table.TransformColumnTypes(#"Removed Duplicates",{{"popularity", Int64.Type},{"revenue", Int64.Type}, {"runtime", Int64.Type}, {"release\_date", type text}, {"budget", Int64.Type},{"vote\_count", Int64.Type}, {"vote\_average", Int64.Type}}),

#"Inserted Parsed JSON" = Table.AddColumn(#"Changed Type", "JSON", each Json.Document([genres])),

#"Expanded JSON" = Table.ExpandListColumn(#"Inserted Parsed JSON", "JSON"),

#"Inserted Parsed JSON1" = Table.AddColumn(#"Expanded JSON", "JSON.1", each Json.Document([genres])),

#"Removed Columns" = Table.RemoveColumns(#"Inserted Parsed JSON1",{"JSON.1"}),

#"Expanded JSON1" = Table.ExpandRecordColumn(#"Removed Columns", "JSON", {"id", "name"}, {"JSON.id", "JSON.name"}),

#"Renamed Columns" = Table.RenameColumns(#"Expanded JSON1",{{"JSON.name", "Genre"}, {"JSON.id", "Genre Id"}}),

#"Inserted Parsed JSON2" = Table.AddColumn(#"Renamed Columns", "JSON", each Json.Document([spoken\_languages])),

#"Expanded JSON2" = Table.ExpandListColumn(#"Inserted Parsed JSON2", "JSON"),

#"Expanded JSON3" = Table.ExpandRecordColumn(#"Expanded JSON2", "JSON", {"iso\_639\_1", "name"}, {"JSON.iso\_639\_1", "JSON.name"}),

#"Renamed Columns1" = Table.RenameColumns(#"Expanded JSON3",{{"JSON.name", "spoken language"}}),

#"Removed Columns1" = Table.RemoveColumns(#"Renamed Columns1",{"genres"}),

#"Inserted Parsed JSON3" = Table.AddColumn(#"Removed Columns1", "JSON", each Json.Document([production\_companies])),

#"Expanded JSON4" = Table.ExpandListColumn(#"Inserted Parsed JSON3", "JSON"),

#"Expanded JSON5" = Table.ExpandRecordColumn(#"Expanded JSON4", "JSON", {"name", "id"}, {"JSON.name", "JSON.id"}),

#"Renamed Columns2" = Table.RenameColumns(#"Expanded JSON5",{{"JSON.name", "production name"}, {"JSON.id", "production id"}}),

#"Removed Columns2" = Table.RemoveColumns(#"Renamed Columns2",{"spoken\_languages"}),

#"Inserted Parsed JSON4" = Table.AddColumn(#"Removed Columns2", "JSON", each Json.Document([production\_countries])),

#"Renamed Columns3" = Table.RenameColumns(#"Inserted Parsed JSON4",{{"production name", "production company name"}}),

#"Expanded JSON6" = Table.ExpandListColumn(#"Renamed Columns3", "JSON"),

#"Expanded JSON7" = Table.ExpandRecordColumn(#"Expanded JSON6", "JSON", {"iso\_3166\_1", "name"}, {"JSON.iso\_3166\_1", "JSON.name"}),

#"Renamed Columns4" = Table.RenameColumns(#"Expanded JSON7",{{"JSON.iso\_3166\_1", "prd country id"}, {"JSON.name", "prd countries"}}),

#"Removed Columns3" = Table.RemoveColumns(#"Renamed Columns4",{"production\_countries", "production\_companies"}),

#"Inserted Parsed JSON5" = Table.AddColumn(#"Removed Columns3", "JSON", each Json.Document([keywords])),

#"Expanded JSON8" = Table.ExpandListColumn(#"Inserted Parsed JSON5", "JSON"),

#"Expanded JSON9" = Table.ExpandRecordColumn(#"Expanded JSON8", "JSON", {"id", "name"}, {"JSON.id", "JSON.name"}),

#"Renamed Columns5" = Table.RenameColumns(#"Expanded JSON9",{{"JSON.id", "kyword id"}, {"JSON.name", "keyword name"}}),

#"Removed Columns4" = Table.RemoveColumns(#"Renamed Columns5",{"keywords"})

in

#"Removed Columns4"