Extract:

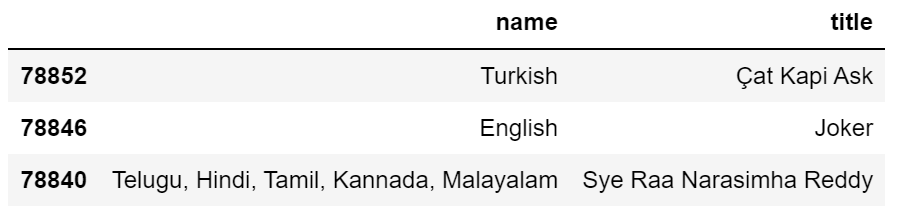
The original source data was already in four CSV files retrieved from Kaggle and World Data websites.

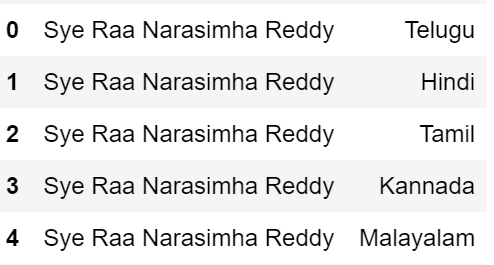
Transform:

**Country-Language Tables:**

Source: IMBD Movie Excel

1. Original excel was loaded as a Pandas Dataframe.
2. Data was filtered to get only movies and the rest of columns from 2000 to 2019.
3. Filtered DataFrame was cleaned to show only desirable columns.
4. Languages and Countries were columns with comma separated data, so “split” and “stack” pandas’ functions were needed to define a new row every time a cell contains more than one country or language separated by commas, as follow:





1. Columns renaming process to match same database columns names.
2. lang\_id and country\_id serial id’s generation.
3. Reorganizing columns order.

**“lang\_movie” and “movie\_country” Tables:**

Source: IMBD Movie Excel

1. Join each normalized languages and country tables (without comma separation) with the “Movie Table” that contains “movie\_id”. Then columns were renamed.
2. A second Join between the above merged tables (countr\_movie\_merge, lang\_movies\_id\_merge) with their corresponding country\_table and language\_table.

**Rating Table**

Source: IMBD Movie Excel (Excel), MoviesOnStreamingPlatforms\_updated (CSV)

1. Original Movie as Dataframe (movie\_table).
2. Filtered desirable columns
3. MoviesOnStreamingPlatforms\_updated loaded as DataFrame (rating\_source\_df)
4. Columns renaming process.
5. Merge between rating\_source\_df and movie\_table
6. Data cleaning of the merged dataframe (filetered and renamed columns)