IMDB/Bollywood Manipulations and Musings

Chase Khan, Greg Mika

The Raw Datasets

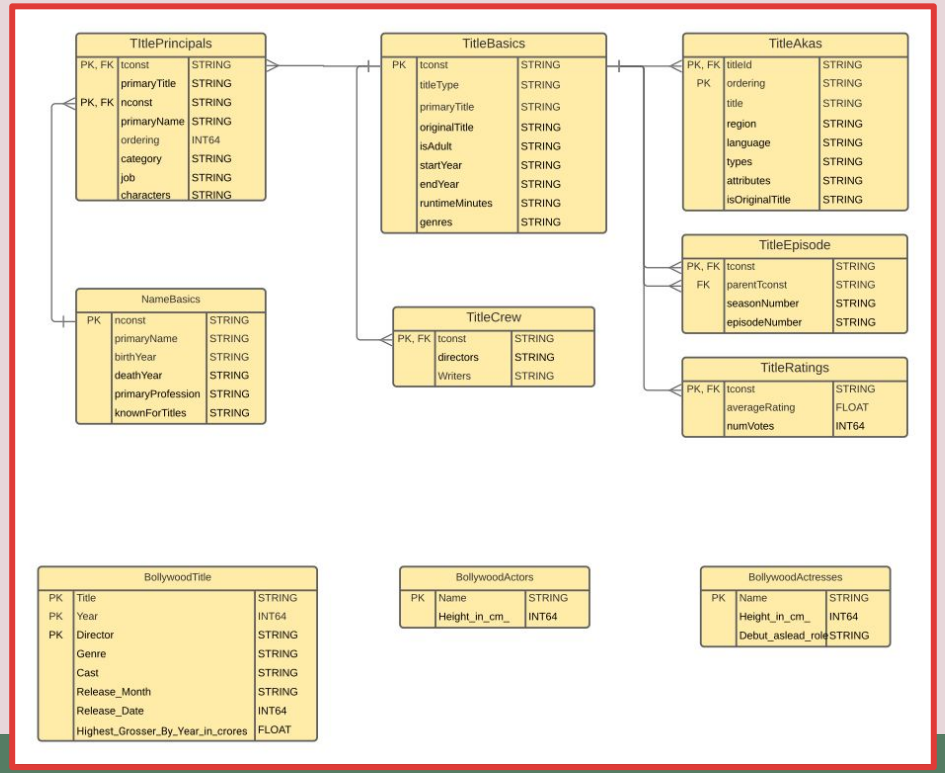
IMDB

- NameBasics
- TitleBasics
- TitleCrew
- TitleEpisode
- TitlePrincipals
- TitleRatings
- TitleAKAs

Bollywood

- Bollywood Actors
- Bollywood Actresses
- Bollywood (Titles)

The Staging Tables



Beam Pipeline Goals

Directors, Writers

Parse a string array

Make a junction table

Join with NameBasics

Bollywood

Create a unique id for the BollywoodTitles Table

Genres (IMDb and Bollywood), Professions

Parse string arrays

Make a child table

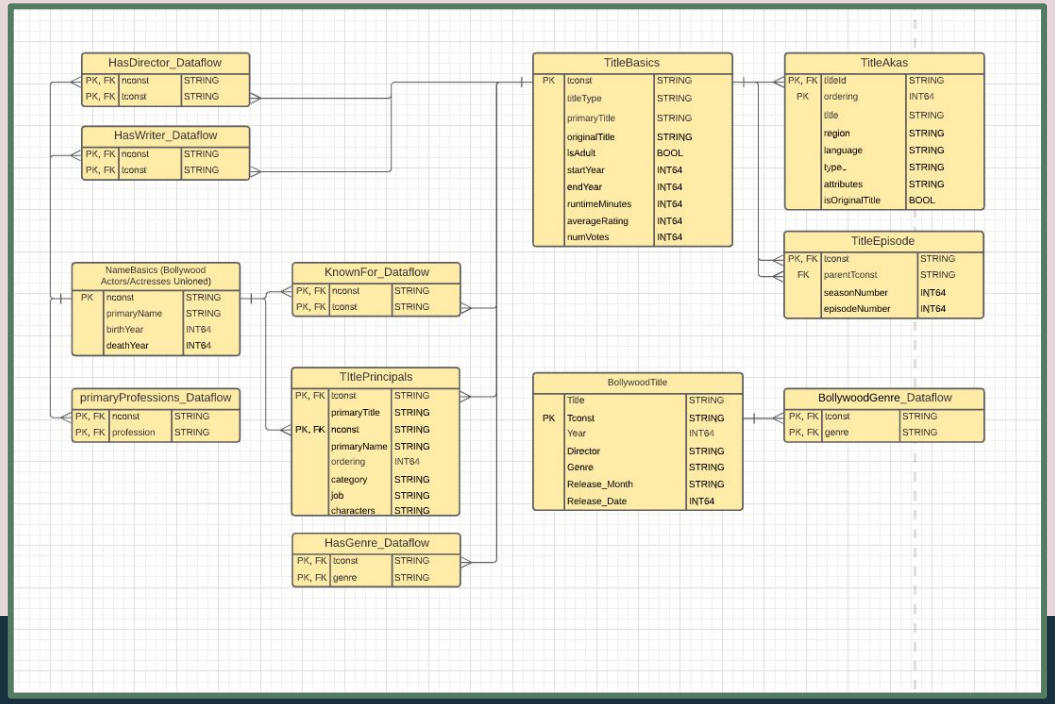
Known For

Parse a string array

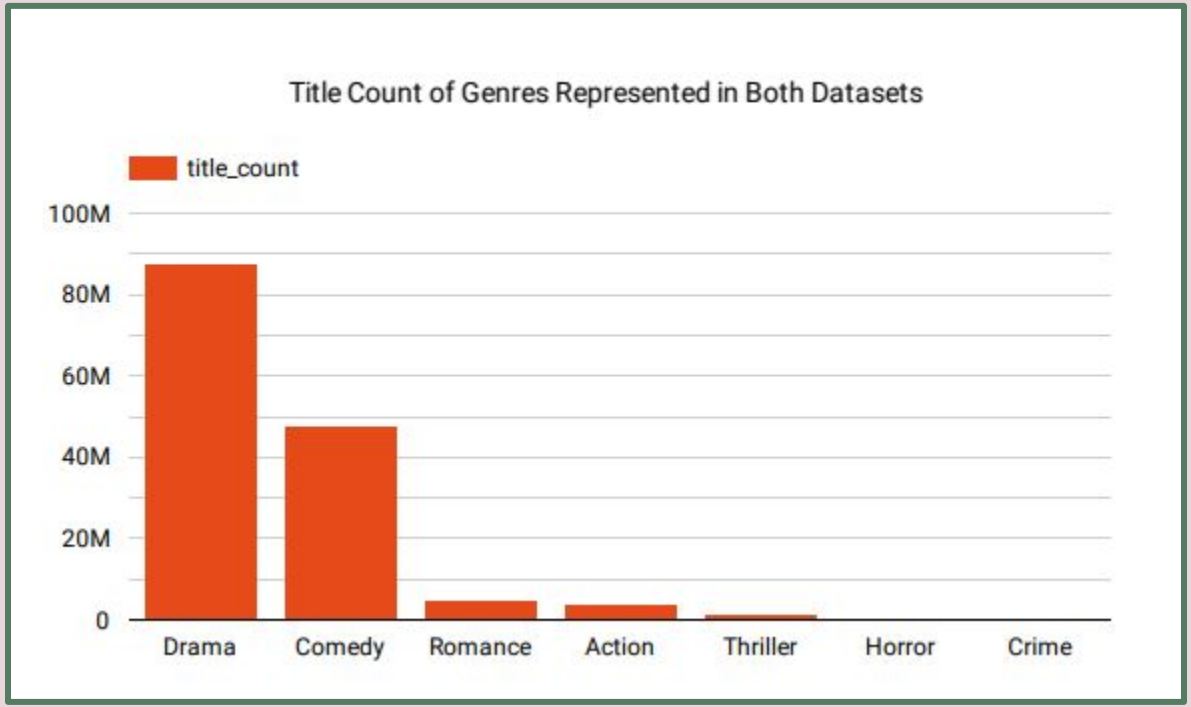
Make a junction table

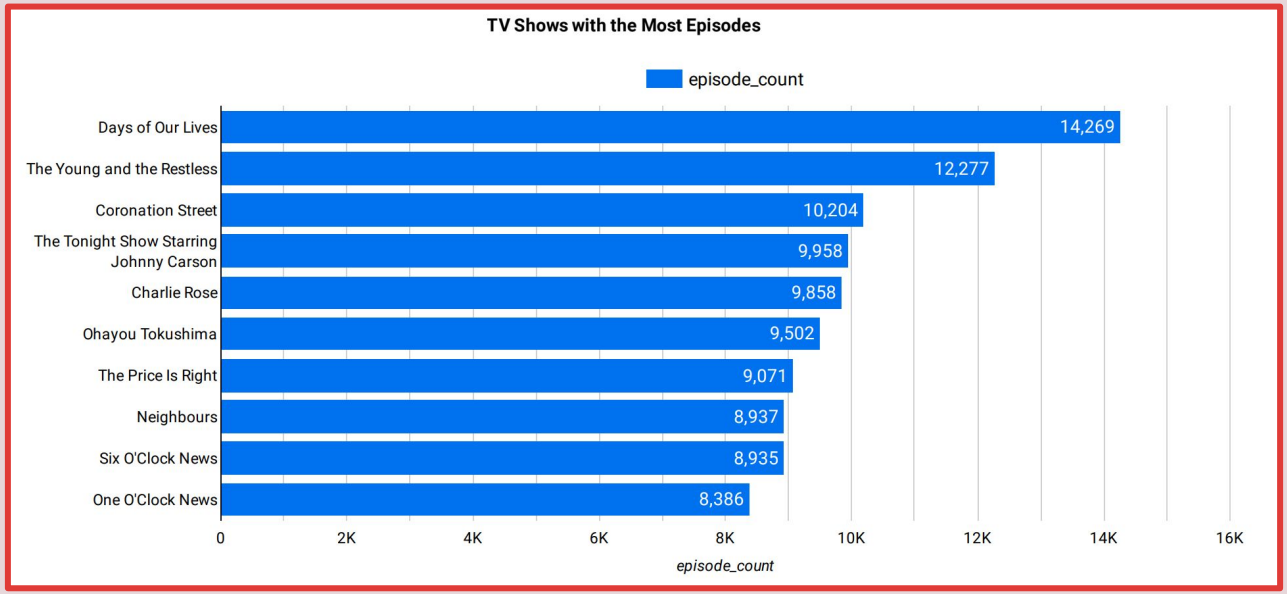
Join with TitleBasics

The Modeled Tables



Cross Dataset Queries


Title Count of Genres Represented in Both Datasets
title_count
100M
80M
60M
40M
20M
0
Drama
Comedy
Romance
Action
Thriller
Horror
Crime


TV Shows with the Most Episodes
episode_count
Days of Our Lives
The Young and the Restless
Coronation Street
The Tonight Show Starring Johnny Carson
Charlie Rose
Ohayou Tokushima
The Price Is Right
Neighbours
Six O'Clock News
One O'Clock News
14,269
12,277
10,204
9,958
9,858
9,502
9,071
8,937
8,935
8,386
0
2K
4K
6K
8K
10K
12K
14K
16K
episode_count

Future Improvements

- Write more queries to compare the most popular genre in each dataset, highest rated genres, percentage of titles that make a up genre, etc.

- Integrate the Bollywood titles table into the IMDb title table
  - Find titles that already exist in the IMDb table
  - Add Bollywood titles not in the IMDb table and give them a primary key of the same format as the IMDb table

- Find another secondary data source to fill in gaps in the Bollywood data.