Data description

I dropped some columns and renamed some.

Our movies.csv now has:

* movieId (previously titleId)
* region
* language
* types
* title (previously primaryTitle)
* isAdult
* year(previously startYear)
* runtime (previously runtimeMinutes)
* genres

The casts.csv now has:

* castId (previously nconst)
* primaryName
* birthYear
* deathYear
* roles (previously primaryProfession)
* movies (previously knownForTitles)

Matteo suggested we break down the genre column (movies.csv) and the roles column (cast.csv)

The unique roles are:

[1] "soundtrack" "actor" "miscellaneous"

[4] "actress" "music\_department" "writer"

[7] "director" "producer" "stunts"

[10] "make\_up\_department" "composer" "assistant\_director"

[13] "camera\_department" "music\_artist" "editor"

[16] "cinematographer" "casting\_director" "script\_department"

[19] "art\_director" "costume\_department" "animation\_department"

[22] "art\_department" "executive" "special\_effects"

[25] "production\_designer" "production\_manager" "editorial\_department"

[28] "sound\_department" "talent\_agent" "casting\_department"

[31] "costume\_designer" "visual\_effects" "location\_management"

[34] "set\_decorator" "transportation\_department" NA

[37] "manager" "publicist" "legal"

[40] "assistant" "podcaster" "production\_department"

[43] "choreographer" "electrical\_department"

The unique genres are:

[1] "Documentary" "Short" "Animation" "Comedy" "Romance" "Sport" "News"

[8] "Drama" "Fantasy" "Horror" "Biography" "Music" "War" "Crime"

[15] "Western" "Family" "Adventure" "Action" "History" "Mystery" "\\N"

[22] "Sci-Fi" "Musical" "Thriller" "Film-Noir" "Talk-Show" "Game-Show" "Reality-TV"

[29] "Adult" NA