Name: Sasha Kenkre
Uniqname: skenkre

Database Design Choices

My original datasource was a boxoffice mojo dataset from Kaggle. There were 26 columns in it, but I chose to narrow that down to 16 columns from the original dataset. Here are some of the choices I made:

- There were 4 columns for actors and 4 columns for genres in the original dataset. To cut down on filtering through a lot of data, I chose to only keep the first column of each as that seemed to be the main column for those categories. This would keep my main movie table cleaner.
- I converted runtime from the original data (hours and minutes) to minutes.
- I selected only data from 2018-2020 as there was too much data otherwise.
- It made sense to have one main ‘movie’ table with four supporting tables that would help facilitate foreign keys.
- One of the main areas where I saw overlap was for people. This was through the director, writer, producer, composer, cinematographer, and actor columns. So I created a table that included the names of all these people, called ‘person’ and used ids within the main table to cut down on redundancy of names.
    - I originally thought I would need a table for the type of role (producer, writer, etc.), however, this proved complicated and confusing when I tried to do that along with the person table. So I scrapped that idea, and kept the separate columns for roles and had foreign keys that referenced the ‘person’ table.
- Since there were many duplicate categories in genre, distributors, and rating, I made tables for each and had the main ‘movie’ table have foreign keys referencing their respective tables.
- I did not use ‘cascade for delete’ when making foreign keys because deleting any one piece of information could affect other unrelated movies, which I felt would create issues in the data.