CSC343 Project - Proposal

1. Domain

Study about factors affecting movie ratings across various rating websites.

2. Dataset

Link	Information relevant to project	Learning will have to do to interpret data	Cleaning up we need
https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset	Mid (IMDb title id), title, year, date, genre, duration, country, language, director, writer, actors, budget, revenue	- special characters in the file should be able to display properly - for an attribute that contains a list of string, we should be able to search inside the list for a certain string	- remove attributes that we don't need - separate attributes to different relations, and then remove tuples with null values
https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+names.csv	Pid (IMDb name id), name, birth year, place of birth	- how to get a list of people born within a certain year range - logically, other Pids should be a subset of this Pid, but how do we deal with missing data in this Pid (same problem for Mid)	
https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+ratings.csv	Mid, IMDb rating		

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+title_principals.csv	Mid, Pid, job type	- how to get all existing values for an attribute and save them for later use (we want to know what job types were included)	
https://www.kaggle.com/stephanerappeneau/350-000-movies-from-themoviedborg	Pid, award, year, outcome		
https://www.kaggle.com/juzershakir/tmdb-movies-dataset	Mid, budget, revenue		- remove tuples with 0 budget or 0 revenue
https://www.kaggle.com/stefanoleone992/filmtv-movies-dataset	filmtv id, title, filmtv rating	- link filmtv id to IMDb id (our Mid)	
https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset?select=rotten_tomatoes_movies.csv	rotten tomatoes id, title, rotten tomatoes rating	- link rotten tomatoes id to IMDb id	

3. Investigative Questions

- Do movies made by more experienced producing teams have a higher chance to get a good rating? For example, do movies directed by renowned and award-winning directors tend to have better ratings?
- Does high financial investment on a movie necessarily increase the rating and is the high gross also linked to a high rating? For example, is it possible for a movie to have a relatively low budget and a low revenue, but have a very good rating?
- Will rating websites have biases towards some genres of movie? For example, do documentaries always get rated higher in RottenTomatoes than in Filmtv?

4. Schema

a. Relational Schema

Movie (<u>Mid</u>, title, Myear, Mdate, country, language, duration, genre)
A tuple in this relation represents a movie. Mid is the IMDb title id of the movie; title is the movie title; Myear is the year when the movie was released; Mdate is the date when the movie was released; country is the movie producing country; language is the main language the movie uses; duration is the movie duration in minutes; genre is the movie genre.

Person (<u>Pid</u>, name, birthDate, birthPlace)
A tuple in this relation represents a person. Pid is the IMDb name id of the person; name is the full name of the person; birthDate is the birthdate of the person; birthPlace is the birthplace of the person.

Award(<u>Pid, awardName, Ayear</u>, outcome)
A tuple in this relation represents an award. Pid is the IMDb name id of the person; awardName is the name of an award for the person. Ayear is the year when the person won the award or got nominated; outcome is whether the person actually won the award or just got nominated.

Job(<u>Mid, Pid, jobCategory</u>)
A tuple in this relation represents a job related to a movie. Mid is the IMDb title id of the movie; Pid is the IMDb name id of the person; jobCategory is the person's specific job in this movie production.

Finance(<u>Mid</u>, Budget, Revenue)
A tuple in this relation represents the finance information of a movie. Mid is the IMDb title id of the movie; Budget is the budget of producing the movie; Revenue is the movie revenue after it was released.

Rating(<u>Mid</u>, IMDb, RottenTomatoes, FilmTV)
A tuple in this relation represents movie ratings of a movie across various rating websites. Mid is the IMDb title id of the movie; IMDb is IMDb rating for the movie; RottenTomatoes is RottenTomatoes rating for the movie; FilmTV is FilmTV rating for the movie.

b. Constraint

Finance[Mid] $\subseteq$ Movie[Mid]
Job[Mid] $\subseteq$ Movie[Mid]
Job[Pid] $\subseteq$ Person[Pid]
Award[Pid] $\subseteq$ Job[Pid]
Rating[Mid] $\subseteq$ Movie[Mid]
Myear and Mdate should not be larger than today's date.

c. Data Dictionary

Attribute	Description	Type	Always be known	Default value	Allowable values
Mid	The IMDb title id of the movie	TEXT	yes	N/A	Natural numbers
title	Title of the movie	TEXT	yes	N/A	Valid movie name
Myear	The year when the movie was released	INT	yes	N/A	From 1894 (the year that the first movie released) to current year
Mdate	The date when the movie was released	date format	yes	N/A	Valid date
country	The producing country	TEXT	yes	N/A	Switzerland, Canada, Japan, Germany, United Kingdom, United States, etc.
language	The main language the movie uses	TEXT	yes	N/A	Spanish, English, Germanic, Chinese, Portuguese, etc.
duration	The movie duration (in minutes)	INT	yes	N/A	Valid duration
genre	The genre of movie	TEXT	yes	N/A	Romance, Biography, Crime, Drama, History, Adventure, Crime, Fantasy, Family, Horror, Comedy, etc.
Budget	The budget of movie	INT	yes	N/A	Natural Numbers
Pid	The unique ID of the person	TEXT	yes	N/A	Natural numbers

Revenue	The revenue of movie	INT	yes	N/A	Natural Numbers
name	Name of the person	TEXT	yes	N/A	Valid person name
birthDate	Birthdate of the person	date format	yes	N/A	Valid date
birthPlace	Birthplace of the person	TEXT	yes	N/A	Valid place
jobCategory	The person's job in certain movie	TEXT	yes	N/A	actor, actress, director, cinematographer, producer, writer, etc.
awardName	The name of an award for a director, actor, actress or others.	TEXT	yes	N/A	The Academy Award for Best Picture, Best Director, Best Actor, Best Actress, etc.
Ayear	The year when the person won the award or got nominated.	INT	yes	N/A	1894-current year
outcome	Whether the person won the award or got nominated.	TEXT	yes	N/A	'Nominated' OR 'Won'
IMDb	IMDb rating for the movie	INT	yes	N/A	0-10
RottenTomatoes	RottenTomatoes rating for the movie	INT	yes	N/A	0-100
FilmTV	FilmTV for the movie	INT	yes	N/A	0-10

d. Justification of Design

- The ‘Movie’ relation contains only basic information about movies, which allows us to keep most tuples from the dataset.
- The ‘Person’ relation contains only basic information about the person, which allows us to keep most tuples from the dataset.
- The ‘Job’ relation connects people and movies, i.e., what the person’s job is in the specific movie.
- The ‘Rating’ relation allows us to compare how the ratings vary in different rating websites.
- We separate ‘Award’ from the ‘Person’ since many movie producing or acting staff did not win/nominate an award.
- We separate ‘Finance’ from the ‘Movie’ since many movies do not have their budget and revenue information available online.