**Codebook ‘ChangingThePicture\_mastersheet.csv’**

ChangingThePicture\_mastersheet.csv contains metadata for 58.444 releases, including films, TV series and documentaries. Collection of these films started with the 617 films in [the Cornell movie dialogues corpus](https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html). The selection was then expanded to include all releases that shared one or more crew members with these 617 films, resulting in a collection of 58.444 releases. For this, we used the publicly available [IMDb data files](https://datasets.imdbws.com/).

Please note that the data set has 102.902 rows. This is because every director has their own row - a film with 4 directors will therefore be represented by 4 rows. This makes it easier to work with director-specific data (such as directorPreviousFilmsCount).

**tconst** (string)

IMDb release ID. Used in IMDb hyperlinks (e.g. <https://www.imdb.com/title/tt0103594>).

**titleType** (string)

Release type, e.g. movie, tvSeries, short, videogame.

**primaryTitle** (string)

Main title for the film.

**releaseYear** (integer)

Year in which the release came out (according to IMDb).

**genres** (list)

Comma-separated list of genres, with a maximum of 3.

**nconst\_director** (string)

IMDb person ID. This column only contains IDs for movie directors. Used in IMDb hyperlinks (e.g. <https://www.imdb.com/name/nm0001008/>).

**imdbRating** (numeric)

Average IMDb score for the film.

**meanPreviousIMDbRatingDirector** (numeric)

Mean of IMDb ratings of previous releases by director.

**directorPreviousFilmsCount** (integer)

Number of releases by the director released before the current release.

**crewCount** (integer)

Amount of crew members, as registered in IMDb. Maximum of 10 per film.

**percentageFemaleCrew** (numeric)

The percentage of female crew members for the film. Expressed by number between 0 and 1.

**Rated** (string)

*Only for films in the Cornell data set.* US age rating of the film (e.g. R, PG-13).

**Writer** (string)

*Only for films in the Cornell data set.* Names of the film writers.

**Actors** (string)

*Only for films in the Cornell data set.* Names of the film actors.

**Plot** (string)

*Only for films in the Cornell data set.* Synopsis of the film plot.

**Language** (string)

*Only for films in the Cornell data set.* Language of the film.

**Country** (string)

*Only for films in the Cornell data set.* Film’s country of origin.

**Awards** (string)

*Only for films in the Cornell data set.* Awards and nominations the film won.

**Metascore** (integer)

*Only for films in the Cornell data set.* Metacritic score of the film.

**BoxOffice** (string)

*Only for films in the Cornell data set.* Box office of the film.

**BT.id** (string)

*Only for films in the Cornell data set.* ID for the film in [the Bechdel Test data base](https://bechdeltest.com/).

**BT.score** (integer)

*Only for films in the Cornell data set.* Bechdel Test score of the film, expressed by a number between 0 and 3. As described on the above-mentioned website: “(1) it has to have at least two women in it, who (2) who talk to each other, about (3) something besides a man.”

**BT.dubious** (logical)

*Only for films in the Cornell data set.* Whether the Bechdel Test score is contested by the community, meaning the score might be unreliable.

**char\_percentage\_female** (numeric)

*Only for films in the Cornell data set.* The percentage of female characters in the film. Expressed by number between 0 and 1.

**dir\_percentage\_female** (numeric)

*Only for films in the Cornell data set.* The percentage of female directors for the film. Expressed by number between 0 and 1.

**nr\_nodes** (integer)

**nr\_edges** (integer)

**density** (numeric)

**triad\_closure** (numeric)

**clustering\_coef** (numeric)

**is\_connected** (logical)

**centralization\_degree** (numeric)

**gender\_assortativity** (numeric)