Tab 1

Initial dataset file options :

### **title.akas.tsv.gz**

titleId (string) - a tconst, an alphanumeric unique identifier of the title

* ordering (integer) – a number to uniquely identify rows for a given titleId
* title (string) – the localized title
* region (string) - the region for this version of the title
* language (string) - the language of the title
* types (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New values may be added in the future without warning
* attributes (array) - Additional terms to describe this alternative title, not enumerated
* isOriginalTitle (boolean) – 0: not original title; 1: original title

### **title.basics.tsv.gz**

* tconst (string) - alphanumeric unique identifier of the title
* titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
* primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
* originalTitle (string) - original title, in the original language
* isAdult (boolean) - 0: non-adult title; 1: adult title
* startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year
* endYear (YYYY) – TV Series end year. '\N' for all other title types
* runtimeMinutes – primary runtime of the title, in minutes
* genres (string array) – includes up to three genres associated with the title

### **title.crew.tsv.gz**

* tconst (string) - alphanumeric unique identifier of the title
* directors (array of nconsts) - director(s) of the given title
* writers (array of nconsts) – writer(s) of the given title

### **title.episode.tsv.gz**

* tconst (string) - alphanumeric identifier of episode
* parentTconst (string) - alphanumeric identifier of the parent TV Series
* seasonNumber (integer) – season number the episode belongs to
* episodeNumber (integer) – episode number of the tconst in the TV series

### **title.principals.tsv.gz**

* tconst (string) - alphanumeric unique identifier of the title
* ordering (integer) – a number to uniquely identify rows for a given titleId
* nconst (string) - alphanumeric unique identifier of the name/person
* category (string) - the category of job that person was in
* job (string) - the specific job title if applicable, else '\N'
* characters (string) - the name of the character played if applicable, else '\N'

### **title.ratings.tsv.gz**

* tconst (string) - alphanumeric unique identifier of the title
* averageRating – weighted average of all the individual user ratings
* numVotes - number of votes the title has received

### **name.basics.tsv.gz**

* nconst (string) - alphanumeric unique identifier of the name/person
* primaryName (string)– name by which the person is most often credited
* birthYear – in YYYY format
* deathYear – in YYYY format if applicable, else '\N'
* primaryProfession (array of strings)– the top-3 professions of the person
* knownForTitles (array of tconsts) – titles the person is known for

From these chose : **title.basics.tsv.gz and title.ratings.tsv.gz [Total records :** 9,283,197 **]**

Dataset too large for excel (1 GB+) so uploaded to [gigasheet.com](http://gigasheet.com) for initial inspection. Found in title.basics :

Rows: 408,548 of 9,283,197 - where genre - ‘\N’

Rows: 9,171,714 of 9,283,197 - where endyear - ‘\N’

**Rows:** 1,253,014 of 9,283,197 where startyear - ‘\N’

**Rows:** 6,360,609 of 9,283,197 where runtime - ‘\N’

Steps:

Removed runtime and endyear and originalTitle (same as primaryTitle in most cases so redundant) columns

maps each nconst ID to its corresponding primaryName from df\_names

Split the directors and writers columns, replace nconst IDs with names, and join them into single strings for each title.

Retain only the directorNames and writerNames columns in df\_crew.

Merge Datasets

titleTypes :

movie, short, tvEpisode, tvMiniSeries, tvMovie, tvPilot,tvSeries, tvShort, tvSpecial, video, videoGame

Steps:

Filtered to keep types : movie, tvMovie

New dataset : 319,084 records