table name	column name	column type	column description	column type	has null values ?	null values %	column value	decision	commentary	table changes
title.akas.tsv.gz	**titleId**	string	a tconst, an alphanumeric unique identifier of the title	string	FALSE	0%	high	leave		There are no duplicates in the table.
	ordering	integer	a number to uniquely identify rows for a given titleId	string	FALSE	0%	medium	leave		
	title	string	the localized title	string	FALSE	0%	high	leave		
	region	string	the region for this version of the title	string	TRUE	23%	high	fillna	Missing regions for original titles are replaced with 'original_region', but only for original titles. About 0.26% of missing values will remain. These rows will need to be removed, as they provide no useful information without a region.	
	language	string	the language of the title	string	TRUE	33%	medium	leave	This column may be useful for our business questions, so we will keep it. There is no suitable value to replace the missing entries.	
	types	array	enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay".	string	TRUE	69%	low	delete	This column has many missing values and is not important for our business questions, so it will be excluded.	
	attributes	array	additional terms to describe this alternative title, not enumerated.	string	TRUE	99%	low	delete	99% of the values in this column are missing, and it is not important for our business questions, so it will be excluded.	
	isOriginalTitle	boolean	0: not original title; 1: original title	string	FALSE	0%	low	delete	After replacing null values with 'original_region' in the region column, this column is no longer needed, as we can identify original titles using region == 'original_region'.	
title.basics.tsv.gz	**tconst**	string	alphanumeric unique identifier of the title	string	FALSE	0%	high	leave		There are no duplicates in the table (a total of 11,153,129 rows).
	titleType	string	the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)	string	FALSE	0%	medium	leave		
	primaryTitle	string	the more popular title / the title used by the filmmakers on promotional materials at the point of release	string	FALSE	0%	high	leave		
	originalTitle	string	original title, in the original language	string	FALSE	0%	medium	leave		
	isAdult	boolean	0: non-adult title; 1: adult title	boolean	FALSE	0%	medium	leave		
	startYear	YYYY	represents the release year of a title. In the case of TV Series, it is the series start year	integer	TRUE	13%	high	leave	This column may be important for business questions. Missing values cannot be filled because the range is large (1874–2031), and choosing a single value could introduce bias into the data.	
	endYear	YYYY	TV Series end year. '\N' for all other title types	integer	TRUE	99%	low	delete	99% of the values are missing, and the column is not important.	
	runtimeMinutes	integer	primary runtime of the title, in minutes	integer	TRUE	68%	high	fillna	This column is important for business questions, so missing values were filled using the median for each titleType group.	
	genres	string array	includes up to three genres associated with the title	string	TRUE	5%	high	leave	This column may be important for business questions. Missing values cannot be filled, as choosing a single value could introduce bias into the data.	
title.crew.tsv.gz	**tconst**	string	alphanumeric unique identifier of the title	string	FALSE	0%	high	leave		There are no duplicates in the table.
	directors	array of nconsts	director(s) of the given title	string	TRUE	14%	high	leave	There are records with neither a director nor a writer. These rows are removed, as they provide no useful information. After removal, 5% of values in the two columns remain missing. This means there are still movies with only a director or only a writer.	
	writers	array of nconsts	writer(s) of the given title	string	TRUE	14%	high	leave		
title.episode.tsv.gz	**tconst**	string	alphanumeric identifier of episode	string	FALSE	0%	high	leave		There are no duplicates in the table (a total of 8,583,561 rows).
	parentTconst	string	alphanumeric identifier of the parent TV Series	string	FALSE	0%	high	leave		
	seasonNumber	integer	season number the episode belongs to	integer	TRUE	20%	medium	fillna	This column is important for business questions. Missing values will be replaced with 1 (assuming there is only 1 season), since the table contains information only about TV series.	
	episodeNumber	integer	episode number of the tconst in the TV series	integer	TRUE	20%	medium	delete	This column is not important for business questions, so it will be removed.	
title.principals.tsv.gz	**tconst**	string	alphanumeric unique identifier of the title	string	FALSE	0%	high	leave		There are no duplicates in the table (a total of 88,603,473 rows).
	ordering	integer	a number to uniquely identify rows for a given titleId	integer	FALSE	0%	high	leave		
	nconst	string	alphanumeric unique identifier of the name/person	string	FALSE	0%	high	leave		
	category	string	the category of job that person was in	string	FALSE	0%	medium	leave		
	job	string	the specific job title if applicable, else '\N'	string	TRUE	81%	low	delete	This column has many missing values (81%) and is not important for our business questions, so it will be removed.	
	characters	string	the name of the character played if applicable, else '\N'	string	TRUE	52%	low	delete	Just over a half of the values are missing, and the column is not relevant, so it will be deleted.	

title.ratings.tsv.gz	**tconst**	string	alphanumeric unique identifier of the title	string	FALSE	0%	high	leave		There are no duplicates in the table.
	averageRating	float	weighted average of all the individual user ratings	double	FALSE	0%	high	leave		
	numVotes	integer	number of votes the title has received	integer	FALSE	0%	high	leave		
name.basics.tsv.gz	**nconst**	string	alphanumeric unique identifier of the name/person	string	FALSE	0%	high	leave		There are no duplicates in the table (a total of 13,884,604 rows; after removing rows with missing values in primaryName, 13,884,554 rows remain).
	primaryName	string	name by which the person is most often credited	string	TRUE	0.00036%	high	leave	Rows with missing values in primaryName also have missing values in primaryProfession and knownForTitles. There are only 50 such rows, so they are removed.	
	birthYear	YYYY	in YYYY format	integer	TRUE	95%	low	delete	This column has 95% missing values and is not important for our business questions.	
	deathYear	YYYY	in YYYY format if applicable, else '\N'	integer	TRUE	98%	low	delete	This column has 98% missing values and is not important for our business questions.	
	primaryProfession	array of strings	the top-3 professions of the person	array of strings	TRUE	19%	high	fillna	This column is important for business questions. Missing values are filled with the mode of the primaryProfession column (after calculation, 'actor' is the mode).	
	knownForTitles	array of tconsts	titles the person is known for	array of strings	TRUE	11%	medium	fillna	This column may be important for business questions. Missing values are filled with an empty list [], since filling with the mode is not as appropriate here as in the previous case.	

IMDb Dataset Details

GitHub Repository: https://github.com/tetianasokolova/big-data-project.git