Data and Information Quality

## ID Project: 3

## Dataset: 1

Luca Mattana – ID Person: 10104067

# Introduction and setup choices

This report outlines the *data preparation* steps undertaken with focus on profiling, assessing, and cleaning the assigned dataset 1: *Movies*. the project pipeline included detailed phases of data quality assessment, data profiling, data wrangling and various cleaning operations such as data transformation, error correction, and deduplication.

The original dataset is composed of 9999 records, 9 columns.

|  |
| --- |
| # Column Non-Null Count Dtype  --- ------ -------------- -----  0 MOVIES 9999 non-null object  1 YEAR 9355 non-null object  2 GENRE 9919 non-null object  3 RATING 8179 non-null float64  4 ONE-LINE 9999 non-null object  5 STARS 9999 non-null object  6 VOTES 8179 non-null object  7 RunTime 7041 non-null float64  8 Gross 460 non-null object |

Considering the original attribute “MOVIES” (titles) there are 6817 unique values and no null values.

Some attribute like YEAR and GROSS contain not homogeneous values (digits + letters + special chars) and they required some actions during data preparation.

For some attributes I have decided to perform a *split*. For example, I have made a split on the original YEAR attribute to manage values containing a range of years like (2010-2020) adding a column with the last year of the range. A split is performed for attribute STARS too to separate the Directors from the list of Stars, originally assembled under the same column.

The *null* values that I can count also with the table above are not a real image of all the null values present in the dataset: some values of the dataset are populated only with’\n’ that can be considered as null equivalent and for the attribute ONE-LINE there is a considerable count of value ‘Add a Plot’ that is equivalent to not a value too. During the data preparation these exceptions are managed and they are explained in detail in the commented code and later in this report.

For the *deduplication* I have decided to not reduce to one title for 2 main reasons

* there could be ad here are for sure movies with same title.
* some episodes of series of TV show with the same title have a lot of populated attributes like the plot, the rating and votes, the different stars and different directors for each episode or season.

To intercept what to deduplicate or not based on these considerations, I have performed multiple checks with *record linkage* with *blocking*. More details in the code and in the specific section *Deduplication* of this report.

# Pipeline implementation

## Data Quality Assessment

The data quality assessment focused on evaluating key dimensions on the original dataset before eany operation:

#### Completeness

#### Accuracy

The accuracy it was evaluated for the original attribute GENRE to check if there were some misspelled word or duplicated genre for each records

#### Consistency

The consistency it was evaluated considering the rule:

*if votes is not null then rate must be not null and vice versa*

All records are consistent on the rule VOTES 🡨🡪 RATING

#### Uniqueness

The uniqueness it was evaluated for the original attribute MOVIES

#### Distinctness

The distinctness for attribute MOVIES is the same of uniqueness cause no null value.  
It is possible to calculate it for any attribute. For example, before any action on dataset I have evaluated it for the attribute GENRE

## Data Profiling

Data profiling was conducted to understand the structure, content, and overall quality of the dataset.

* Schema and Attribute Analysis
* Summary Statistics
* Missing Data Overview and visualization (with *missingno* library)

Key Findings:

* Some columns contained a high proportion of missing values.
* Data formats were inconsistent in certain fields, requiring standardization.

## Data Wrangling

During the data wrangling process of I have transformed and prepared the dataset performing the following tasks:

Cleaning and Splitting attribute YEAR adding a column

Goals of data wrangling: