The popular IMdb dataset in Kaggle is used for data cleansing project.

Implemented in R, see the details below.

1. Meta data

(original dataset)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Field\_name** | **Description** | **Datatype** | **Level and Format** | **Sample entries** |
| color | The color of the movie film | Text(String) | Nominal | Color, Black and White |
| director\_name | The name of the director of the movie | Text(String) | Nominal (First Name Last name separated by space) | Clint Eastwood, Steven Spielberg |
| num\_critic\_for\_reviews | No:of critic reviews | Number (integer) | Ratio | 20, 340 |
| duration | runtime of the movie | Number (integer) | Interval (time in mins) | 120, 118 |
| director\_facebook\_likes | The number of facebook likes received for the director of the movie | Number (integer) | Ratio | 2000, 130 |
| actor\_3\_facebook\_likes | The number of facebook likes received for the actor\_3 of the movie | Number (integer) | Ratio | 2200, 130 |
| actor\_2\_name | The name of actor\_2 co-actor of the movie | Text(String) | Nominal (First Name Last name separated by space) | Monica Calhoun Rai Alexandra |
| actor\_1\_facebook\_likes | The number of facebook likes received for the actor\_1 of the movie | Number (integer) | Ratio | 10,00,393 |
| gross | the gross revenue of the movie | Number (integer) | Ratio (Currency in USD) | 48325649, 25445740 |
| genres | The genres of the movie | Text(String) | Each genre separated by | | Comedy, Comedy|Drama|Romance |
| actor\_1\_name | The name of main actor of the movie | Text(String) | Nominal (First Name Last name) | Robert De Niro, J.K. Simmons |
| movie\_title | The title of the movie | Text(String) | Nominal | The Salon Insomnia Manica |
| num\_voted\_users | The number of voted users | Number (integer) | Ratio | 8560, 34261 |
| cast\_total\_facebook\_likes | The number of facebook likes received for the cast of the movie | Number (integer) | Ratio | 0, 1406 |
| actor\_3\_name | The name of actor\_3(co-actor) of the movie | String | Nominal (First Name Last name separated by spaces) | Anne Hathaway, Ben Mendelsohn |
| facenumber\_in\_poster | The facenumber in the poster | Number (integer) | Nominal | 0,1,2 |
| plot\_keywords | The keywords that represents the plot of the movie | Text(String) | Each keyword separated by | | based on novel, animal name in title|ape abducts a woman|gorilla|island|king kong |
| movie\_imdb\_link | IMBD web link for the movie | Text(String) | Nominal (http link) |  |
| num\_user\_for\_reviews | The number of user reviews | Number (integer) | Ratio | 1, 324 |
| Language | The main language used in the movie | Text(String) | Nominal | English, French |
| Country | The name of the country, some abbreviated | Text(String) | Nominal | USA, UK |
| content\_rating | The content rating for the movie | Text(String) | Nominal | PG-13 , R |
| Budget | The total amount of budget of the movie | Number (integer) in USD | Ratio(Currency in USD) | 237000000 |
| title\_year | The movie release year | Year | Nominal (year in the format YYYY) | 2016, 1916 |
| actor\_2\_facebook\_likes | The number of facebook likes received for the actor\_2 of the movie | Number (integer) | Ratio | 0, 917 |
| imdb\_score | IMBD Score for the movie out of 10 | Number(decimal) | Ratio (xx.y, at most 2 digits before decimal point and one after) | 1.6, 7.1 |
| aspect\_ratio | aspect ratio for the movie film | Number(decimal) | Ratio(xx.yy at most 2 digits before decimal point and 2 after)) | 1.85, 2.35 |
| movie\_facebook\_likes | The number of facebook likes received for the movie | Number (integer) | Ratio | 0,3000 |

Note: Since the use of the dataset is not given or known at this point, an effort is made to retain most records by implementing various imputations based on a ‘movie’ centred context.

**Cleansing done using RStudio.**

Columns are rearranged and renamed for better organization as below under answer for (b)

Removed columns

1) num\_voted\_users as seem not very useful as imdb\_score is present

2) actor\_1\_facebook\_likes, actor\_2\_facebook\_likes, actor\_3\_facebook\_likes as this information is redundant for the actor and cast\_total\_facebook\_likes present better information with respect to the movie

3) facenumber\_in\_poster – seems not much relevant in the movie dataset context.

New columns

1) sl. Number: the serial number of rows in the dataset

2) actors: A new column is created by concatenating data in actor\_1\_name, actor\_1\_name, actor\_3\_name separated by | and these 3 original columns are removed

b)

Duplicates are removed from the dataset.

> original dataset dimensions

5043 28

> cleansed dataset dimensions

4992 21

|  |  |  |  |
| --- | --- | --- | --- |
| **Field\_Name** | **Has Invalid data?** | **If yes, why?** | **Correction** |
| sl number | No |  | <This is the Sl.no of records in the dataset> |
| movie\_title | Yes | Trailing whitespaces  Presence of Invalid & UTF-8 characters | Removed trailing whitespaces  Removed incorrect UTF character at the end of the string, Replaced other UTF-8 to ASCII |
| movie\_release\_year | Yes | Missing Values | NA fields are replaced by default value 1888.  Why 1888? The year on which first movie is released  *Default value 1888* |
| movie\_runtime | Yes | Outliers  Missing values | Outliers are limited to min of 10 and max of 300.  NA are replaced with median value |
| language | Yes | Value not present | Missing values are from the USA. So, replaced them with English |
| country | Yes | Missing Values | Dropped the rows because, upon looking at the rows, these records don’t seem to have enough information in other columns also. |
| color | Yes | Leading whitespaces  Missing Values | Trimmed leading whitespaces  Replaced with Color, because the year of the movies released is after 1950 when the color movies become predominant. |
| budget | Yes | Outliers  Missing values | Outlier dropped  Missing values are replaced with 0  *Default value 0* |
| gross | Yes | Missing Values | Replaced with 0  *Default value 0* |
| genres | No |  |  |
| content\_rating | Yes | Missing Values | Replaced with 'Not Rated' |
| director\_name | Yes | Missing Values  UTF-8 characters | Replaced with 'Unknown'  Replaced UTF-8 to ASCII |
| actors | Yes | Missing values  UTF-8 characters | This field is a concatenation of 3 actor name fields and separated by |.  Missing values are replaced with 'Unknown'  Replaced other UTF-8 to ASCII |
| aspect\_ratio | Yes | Outliers  Missing values | Outliers are replaced by 1.6  Missing values are replaced by 1.85(most frequent) |
| plot\_keywords | No |  |  |
| num\_critic\_reviews | Yes | Missing Values | Replaced with 0  *Default value 0* |
| num\_user\_reviews | Yes | Missing Values | Replaced with 0  *Default value 0* |
| imdb\_score | No |  |  |
| movie\_facebook\_likes | No |  |  |
| cast\_total\_facebook\_likes | No |  |  |
| director\_facebook\_likes | Yes | Missing Values | Replaced with 0  *Default value 0* |
| movie\_imdb\_link | No |  |  |

Finally, dataset is again checked for duplicates and NA values. None present.