1. **Datasets Introduction**

Since the movie sector is highly popular and essentially digital, access to public movie data is as easy as searching through a search engine. Kaggle has a series of movies dataset so we decided to select our data from there, overriding the need for a web scraping process. We use two main resources for our Data Mining project:

**1.1. The Movies Dataset (MD):** The dataset provides a set of tables with an entity-relationship feature so that we can combine relevant information from different tables. The tables include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

* Location: Kaggle
* Method of acquisition: Simple download
* Problems encountered: The Movies Dataset’s revenue information represents the global revenue which may be harder to predict since many factors may affect the movie’s popularity - like culture, sense of humor, perception of romance etc.
* Resolution: We needed to find another resource that includes only-US-revenue of movies. The Movie Industry Dataset had the necessary information, so we acquired it as our second dataset.

**1.2. Movie Industry Dataset (MI):** The dataset includes just one .CSV file with 6820 movies’ budget and revenue information.

* Location: Kaggle
* Method of acquisition: Simple download
* Problems encountered: Some of the data points did not have budget or revenue information.
* Resolution: We needed to wipe off the data points which did not have budget or revenue information. We ended up with 4638 records after filtering out 2182 movies.

1. **Datasets Description**

The Movies Database have two tables: "metadata" and "keywords". The first table includes basic information about the movie while the keywords table consists of related keywords for each movie. They can be linked together using "id" column.

Movie Industry dataset have one table: "movies". This table includes all the relevant information about the movies, including our target, profitability.

Number of records and fields in each table is:

|  |  |  |  |
| --- | --- | --- | --- |
| Database | Table Name | Records | Fields |
| Movies Database (MD) | Metadata | 45466 | 11 |
| Movies Database (MD) | Keywords | 46419 | 2 |
| Movies Industry (MI) | Movies | 6820 | 18 |

**2.1. Type of data.**

Our data used for the project is **a mix of structured and semi-structured data.**

While MI dataset is completely structured, MD dataset includes data where the values are either structured or semi-structured, particularly in JSON format. We convert the semi-structured data into structured data using Python’s *json* library.

**2.2. Type of datasets.**

MI dataset is **a single-file tabular dataset**, while the MD dataset is **a relational database** data. We converted MI such that it is a part of the MD dataset’s schema. We used movies’ titles for joining since MI data did not have movie IDs as MD data did.

**2.2. Types of instances of our datasets.**

**MD - Metadata Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **field** | **data\_format** | **is\_descriptive** | **is\_target** | **data\_attribute** |
| **adult** | object | 1 | 0 | categorical -> nominal -> binary |
| **id** | object | 0 | 0 | categorical -> nominal |
| **imdb\_id** | object | 0 | 0 | categorical -> nominal |
| **original\_title** | object | 1 | 0 | categorical -> nominal |
| **overview** | object | 1 | 0 | categorical -> nominal |
| **popularity** | float64 | 0 | 0 | continuous |
| **tagline** | object | 1 | 0 | categorical -> nominal |
| **title** | object | 1 | 0 | categorical -> nominal |
| **genres\_edited** | object | 1 | 0 | categorical -> nominal |
| **spoken\_languages\_edited** | object | 1 | 0 | categorical -> nominal |
| **production\_countries\_edited** | object | 1 | 0 | categorical -> nominal |

**MD - Keywords Dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **field** | **data\_format** | **is\_descriptive** | **is\_target** | **data\_attribute** |
| **id** | int64 | 0 | 0 | categorical -> nominal |
| **keywords\_edited** | object | 1 | 0 | categorical -> nominal |

**MI – Movies Dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **field** | **data\_format** | **is\_descriptive** | **is\_target** | **data\_attribute** | **example** |
| **budget** | int64 | 0 | 0 | continuous | 8000000 |
| **company** | object | 0 | 0 | categorical -> nominal | Columbia Pictures Corporation |
| **country** | object | 1 | 0 | categorical -> nominal | USA |
| **director** | object | 1 | 0 | categorical -> nominal | Rob Reiner |
| **genre** | object | 1 | 0 | categorical -> nominal | Adventure |
| **gross** | int64 | 0 | 0 | continuous | 52287414 |
| **name** | object | 1 | 0 | categorical -> nominal | Stand by Me |
| **rating** | object | 0 | 0 | categorical -> nominal | R |
| **released** | object | 1 | 0 | continuous -> date | 1986-08-22 00:00:00 |
| **runtime** | int64 | 1 | 0 | continuous | 89 |
| **score** | float64 | 0 | 0 | continuous | 8.1 |
| **star** | object | 0 | 0 | categorical -> nominal | Wil Wheaton |
| **votes** | int64 | 0 | 0 | discrete | 299174 |
| **writer** | object | 1 | 0 | categorical -> nominal | Stephen King |
| **year** | int64 | 1 | 0 | categorical -> ordinal | 1986 |
| **isprofit** | int64 | 0 | 1 | categorical -> nominal -> binary | 1 |
| **profitability\_ratio** | float64 | 0 | 1 | continuous -> ratio-scaled | 5.53592675 |
| **profitability\_ratio\_bucket** | int64 | 0 | 0 | continuous -> ratio-scaled | 550 |