# Movies Dataset - Project Report

Julián Ferreira
Tiago Antunes
up202202340@fe.up.pt
up201805327@edu.fc.up.pt
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal

## ABSTRACT

In the pursuit of knowledge, data is a collection of discrete values that convey information, describing quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted. This report intends to document the process of extraction a dataset and build a pipeline with data refinement and data analysis on that same dataset.

## KEYWORDS

datasets, data, movies

## 1 INTRODUCTION

Nowadays all information can be turned into data and consequently stored in a database. From the user's point of view, the large amount of data can sometimes be confusing and difficult to understand.

The goal of this report is to deal with a dataset in a way that makes it easy to visualize. For this purpose, we selected a dataset of movies.

## 2 DATASET

### 2.1 Dataset Choice

The chosen dataset was not the first option we thought of. The initial idea was to work with sports data, such as football statistics, but then we realised that it was quite hard to find this type of dataset with rich text, so we ended up changing our minds. The next option is the one we ended up working with, a movies dataset.

First of all, we thought of IMDB. The problem we had with this is that it was not possible to get a free API license, and the free datasets that were available did not have the text we were looking for. After this, we started searching in Kaggle [3]. We found a lot of different options but they were not exactly what we were looking for. But finally we found the one we are working with.

This is a dataset with around 35.000 movies from Wikipedia, which is quite a trustable source. We searched some extra information and we saw it is updated regularly and it is valued in 8.82 in usability by Kaggle [2].

### 2.2 Dataset Content

We have a unique dataset with 34.893 rows that contain the following information:

- `Release year`: The year in which the movie was released
- `Title`: The title of the movie in English
- `Origin/Ethnicity`: The origin of the movie
- `Director`: The name, or names, of the director of the movie
- `Cast`: The names of the main actors
- `Genre`: The genre the movie is qualified as
- `Wiki page`: The link to the page of the Wikipedia of that movie
- `Plot`: A brief summary of the movie

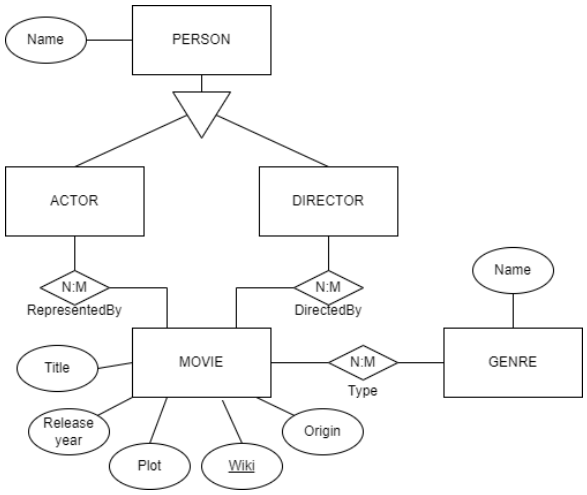


**Figure 1: UML of dataset created on Diagrams.net [1]**

### 2.3 Data Quality and Source

Kaggle is a well known site for data analysts and the dataset is very documentated. It has a ranking of 8.82 in usability so we considered that it is a trustable source. To analyse the quality of the data we did a random search of some movies in the internet to check that our set was correct.

A problem we do have is that there is some missing data. This results in having some empty fields or with the word "unknown". We considered that it is not a huge problem for our work because in some cases it is irrelevant, and also, it is normal that some information is missing in such old movies.

## 3 PIPELINE

Our Pipeline is built almost entirely on python scripts. With the help of *pandas* library we handle and manipulate the data using simple scripts to clean and organize the data.

We didn't put our dataset in a SQL database system and the main reason was that our dataset only contains one file (one table) and wouldn't justify as we don't need to join multiple tables. We chose to keep it simple and our python scripts just handle the dataset perfectly.

,
,



### 3.1 Data Refinement

When it comes to data refinement, the first thing we did was convert the `Genre` column to a list of genres. Imagine three rows of the original dataset:

- `Id`: 1 | ... | `Genre`: drama
- `Id`: 2 | ... | `Genre`: romantic
- `Id`: 3 | ... | `Genre`: drama romantic

If we wanted to know how many different genres are in the example above, the return would be 3 and not 2 as it should be. So we created a little script on python that transform the example above into:

- `Id`: 1 | ... | `Genre`: [drama]
- `Id`: 2 | ... | `Genre`: [romantic]
- `Id`: 3 | ... | `Genre`: [drama, romantic]

This script is not 100% efficacious and it has some flaws. The main idea of this script is to split the Genre string every time it founds a white space. South African, James Bond, Warner Bros are examples of genres with two words and the script sees them as two different genres.

The second and last thing we did about data refinement was to remove the origin movie with less than ten movies. In our data set we had two origins - Assamese and Maldivian - and both had less than ten films. So we removed exactly eleven movies (nine Assamese and two Maldivian) from our dataset.

### 3.2 Data Analysis

In order to obtain more information about the dataset we have developed a python script with several different functions. Each of these try to extract some characterization of the data we are handling and represent the results of the exploration with diagrams or relevant data.

These are the following analysis we carried out:

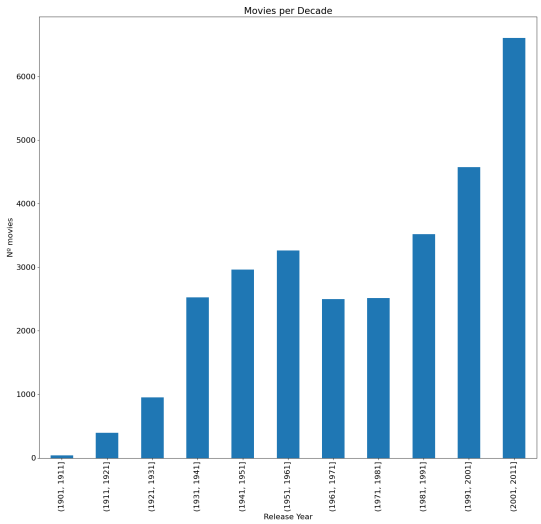


**Figure 2: Movies produced per decade**

Figure 2 represents the number of movies that have been produced in each decade. It is easy to tell that our dataset has the most movies in the 2000s decade, which is quite normal because these are the most recent. Another thing that we could appreciate is that the dataset has more or less the same number of movies from the 1930s to the 1980s because the objective of this dataset is to have information of older movies as well.

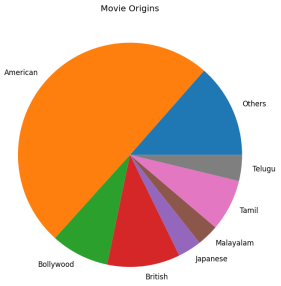


**Figure 3: Movies grouped by origins**

Figure 3 represents the different movies grouped by their origins. In order to represent this we decided to set a minimum number of movies for that origin to appear in the resulting diagram so that it could be easier to understand it. All this genres that did not reach the minimum number are classified as "Others".

About the results obtained, it was easy to predict that most of the movies are from American origin but what was unexpected is that some other ethnicities that are not so known or common also have a high percentage such as Tamil or Bollywood. In conclusion, we can establish that this dataset's objective is to have information with high diversity, even though it might be of not so known movies.

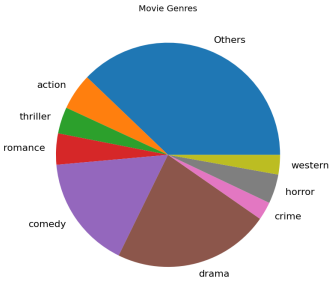


**Figure 4: Movies grouped by genres**

Figure 4 is the result of grouping the movies by their genres, same way as figure 3. For this we did exactly the same as before, establish a minimum number of movies for a genre to appear in the diagram. By analysing the result it is easy to see that the genres in the dataset are really specific because the "Others" genre is the biggest in the chart, which means there are a lot of movies with

specific genres. We can also tell that the most common genres are comedy and drama by far difference from the rest of the genres.

| Keywords (length>5) | |
|---|---|
| Keyword | Frequency |
| before | 13998 |
| father | 13055 |
| police | 12166 |
| family | 11260 |
| becomes | 10254 |
| decides | 10140 |
| through | 9808 |
| himself | 9447 |
| However | 9440 |
| mother | 9057 |

The table above represents the keywords and their frequency of the plot of the movies in our dataset. At first, there was not a minimum length for these keywords but the result was irrelevant because all the words we found were very simple articles which are repeated a lot. So we decided to increase the length and after a couple of different tries we decided to set it as words with more than 5 characters. By doing this, we found much more conclusive results.

The 10 most frequent words in the plots represent the most common themes or expressions in order to do a summary of the movie. It is interesting that three of these words are related with family, which gives us the idea that most of the plots are related to it. Also the word police is very self-explanatory because it represents the high number of movies related with crime in our dataset.

## 3.3 Pipeline Conclusion

The scheme below sumarizes how we obtained, from our initial dataset, multiple plots with some relevant information about the dataset.

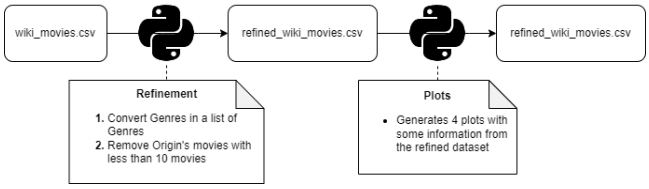


**Figure 5: UML of the pipeline created on Diagrams.net**

# 4 CONCLUSION

We can conclude that our dataset is very rich in terms of diversity, including lots of different types of movies from different ages, but still, the plots of these movies have many related words, which gives us the idea of these having similar stories.

# REFERENCES

[1] Diagrams.net. 2000. Flowchart Maker Online Diagram Software. Retrieved October, 2022 from https://app.diagrams.net/
[2] JUSTINR. 2018. Kaggle: Wikipedia Movie Plots. Retrieved September, 2022 from https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots
[3] Kaggle. 2010. Kaggle: Your Machine Learning and Data Science Community. Retrieved October, 2022 from https://www.kaggle.com/