Introduction to Machine Learning and Data Mining
Danmarks Tekniske Universitet
Fall 2023
Group 50

Report 1

Name	Section 1	Section 2	Section 3	Section 4	Section 5
Csaba Hell - s232456	30%	40%	33.3%	30%	33.3%
Gabriel Lanaro - s233541	40%	30%	33.3%	30%	33.3%
Gabriele Turetta - s233124	30%	30%	33.3%	40%	33.3%

Table 1: Contribution of participants in each section

Contents

1 Description of the dataset

The dataset in question primarily revolves around the world of movies, offering a collection of 626 entries starting from the 1980s to the present day, that can be utilized for various analytical purposes and for conducting predictive analyses. This dataset should provide us valuable insights into the dynamics of the film industry, enabling us to explore trends and patterns among key movie attributes.
The dataset used in this project includes the following key features:

- **MovieID**: A unique identifier for each movie.
- **Title**: The title of the movie. This is a feature that is used solely for identification purposes but is not involved in the calculations or specific analyses of our model.
- **MPAA_Rating**: The Motion Picture Association of America (MPAA) rating of the movie (G – General Audiences, PG – Parental Guidance Suggested, PG-13 – Parents Strongly Cautioned, R – Restricted).
- **Budget**: The budget allocated for producing the movie.
- **Gross**: The total box office gross revenue generated by the movie.
- **Release_Date**: The date when the movie was released.
- **Genre**: The genre or category to which the movie belongs (e.g., Action, Comedy, Drama, etc.).
- **Runtime**: The duration of the movie in minutes.
- **Rating**: The average rating from 0 to 10 given to the movie.
- **Rating_Count**: The number of user ratings received by the movie.

The initial dataset was uploaded in 2020 in data.world, a web portal that offers a wide range of datasets to consult and download. It was originally crafted for educational purposes by Professor James Gaskin, an expert in Information Systems Management at Brigham Young University.
Professor Gaskin employed this dataset to clarify the concepts of relational databases to his students. Importantly, the original dataset is structured as a relational database, which also includes supplementary character and actor datasets. Nevertheless, we intentionally opted to exclude these two additional datasets. We reasoned that the primary dataset provided us with ample information, and merging the additional datasets would have introduced a notable redundancy issue into the final dataset. A picture of the original dataset, including movies, characters and actors, is visible in Figure 8 in the Appendix.
Despite the abundance of information within this dataset, there is no prior analysis or comprehensive study available. Hence, the dataset was mainly used for teaching purposes by Professor Gaskin.

In the context of this film dataset, our goal is to gain insights and make predictions related to various movie attributes. We aim to understand the factors influencing a movie's performance and categorize movies based on different criteria.
More specifically, regarding classification, with this project, we hope to predict

- The *genre* of a movie
- The *MPAA_rating* of a movie

As for regression, our main objectives are to predict:

- The *gross* of a movie
- The *rating* of a movie

To achieve these objectives, a preliminary transformation of some nominal features was necessary to work with them. Specifically, the features *MPAA_rating* and *genre* were converted into numerical values using the label encoding technique, assigning a unique integer to each category.

2 Detailed explanation of the attributes of the data

2.1 Type of attributes

In this section, we delve into the types of attributes within the dataset, classifying them according to their characteristics and measurement scales, as detailed in Table 2.

Movie		
Title	Discrete	Nominal
MPAA rating	Discrete	Nominal
Budget	Continuous	Ratio
Gross	Continuous	Ratio
Release date	Discrete	Interval
Genre	Discrete	Nominal
Runtime	Discrete	Ratio
Rating	Continuous	Ordinal
Rating count	Discrete	Ratio

Table 2: Data types of attributes

2.2 Data issues

The dataset is devoid of any missing values, but it did encounter two instances of data anomalies. These issues were in the *budget* and *gross* attributes, where non-numeric entries such as *x million* were present. Additionally, the *runtime* attribute included mixed units and formatting problems, with one entry containing both *m, h* letters and numerical values.
To address these anomalies, we chose a manual intervention approach. Since only two records were affected, this method proved to be an efficient way to resolve the issues. We manually adjusted the *budget* and *gross* attributes, reformatting them to the correct numerical representation, and standardized the *runtime* attribute for consistency.

2.3 Basic summary statistics of the attributes

Basics statistics

- The number of observations: 626
- The number of variables: 8

Summary for numeric attributes

In this examination, we begin by considering four fundamental metrics: mean, median, minimum, and maximum values. These metrics help us understand different aspects of how the data's typical values and extreme values are distributed within the dataset, giving us a comprehensive view of its overall distribution.

Attribute	Mean	Median	Minimum	Maximum
Budget	94 314 721 $	80 000 000 $	60 000 $	400 000 000 $
Gross	435 227 474 $	345 989 886 $	106 457 $	2 796 000 000 $
Release date	2004.12.29.	2004.12.29.	1989.04.21.	2021.12.17.
Runtime	119 minutes	117 minutes	79 minutes	201 minutes
Rating	6.95	7	4.1	9
Rating count	342 223	252 223	242	2 127 228

Table 3: Mean, median, minimum and maximum values for numeric attributes

In Table 3 we provide a summary of the metrics mentioned before, while in Table 4 we provide additional statistical insights for the same attributes as those presented in Table 3, except *Release date* since it is not suitable for measures of spread or variability. Specifically, we focus on two key metrics, namely the *Standard Deviation* and the *Interquartile Range (IQR)* to offer a deeper understanding of the variability within the dataset.

- **Standard Deviation:** This measurement indicates how much the values for each attribute tend to vary from their average.
- **Interquartile Range (IQR):** This range shows where the middle 50% of the data falls. It's another way to understand the spread of data without being affected by extreme values.

These measurements emphasize significant diversity among the attributes under analysis. Specifically, attributes such as *Gross, Budget,* and *Rating count* demonstrate considerably larger standard deviations, implying substantial variability and the occurrence of possible outliers in the datasets. In contrast, *Runtime* and *Rating* display relatively narrower *Interquartile Ranges (IQR)*, suggesting a more concentrated distribution of values within their middle 50% range.

Attribute	Standard Deviation	Interquartile Range (IQR)
Budget	67 584 298 $	102 000 000 $
Gross	337 969 900 $	355 330 557 $
Runtime	22.38 minutes	31 minutes
Rating	0.87	1.2
Rating count	306 960	313 604

Table 4: Standard deviation and interquartile range for numeric attributes

Summary for categorical attributes

In this section, we delve into the categorical attributes within the dataset, focusing on three key metrics: the Number of Categories, the Mode, and the Frequency of the Mode in Table 5. These metrics collectively provide a comprehensive view of the distribution and prevalence of categories within each attribute.

- **Number of Categories:** This metric shows the diversity of categories within each attribute, representing the degree of classification.
- **Mode:** The mode represents the most frequently occurring category within the attribute, indicating the prevalent category.

- **Frequency of the Mode:** This metric quantifies the prevalence of the mode, offering insights into the dominance of a particular category.

Attribute	Number of Categories	Mode	Frequency of the Mode
MPAA rating	4	G	296 (47.28%)
Genre	17	Action	114 (18.21%)

Table 5: Summary for Categorial Attributtes

- **MPAA Rating:** The attribute comprises four distinct categories, with *G* being the most frequent, appearing 296 times in the dataset and constituting approximately 47.28% of all instances (see Figure 11 in the Appendix), signifying its prevalence in age-appropriateness classification for movies.
- **Genre:** This attribute, encompasses 17 distinct categories, with *Action* emerging as the most prevalent genre category, appearing 114 times and constituting approximately 18.21% of all instances (see Figure 12 in the Appendix), highlighting its dominance among genre classifications.

3 Data visualizations and PCA

3.1 Outliers

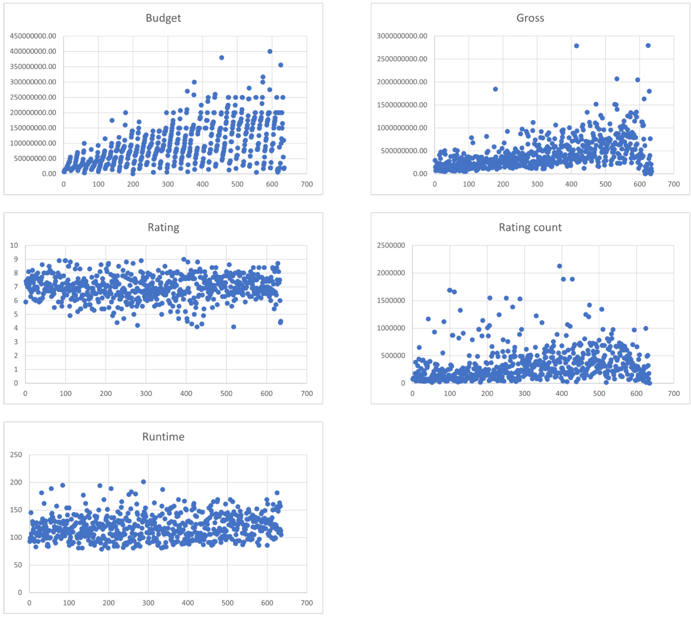

Figure 1: The plots of the features taken into consideration.

As a preliminary step, the following features of the dataset were plotted to observe potential outliers: *budget*, *gross*, *rating*, *rating*, and *rating_ count*. The numerical values of the other features have no intrinsic significance for the detection of outliers and were therefore excluded.

This preliminary analysis allowed us to determine that the data does not show instances that we will consider outliers. Consequently, it was decided to regard the few data points that deviate most from the mean as valid and suitable for PCA. As depicted in Figure 1, where each blue point represents an instance of a film, the

data points that deviate most from the mean, particularly noticeable in the *gross* and *budget* plots, represent some exceptions in the film industry, such as *Avengers: Endgame* in 2019 and *Avatar* in 2009.
We opted not to classify these blockbusters as outliers because they are not merely random exceptions but rather reflect significant trends and important changes in the film industry. These films exemplify the adoption of advanced technologies and the creation of successful franchises, factors that we believe will continue to influence the film market significantly in the near future.

3.2 Normal distributions in the data

About the normal distribution of the data, the features that could present such a distribution are *budget, gross, runtime, rating* and *rating_ count.* We then plotted the frequency histograms of these features, overlaying a Gaussian curve with mean and standard deviation equal to the mean and standard deviation of the analysed feature.

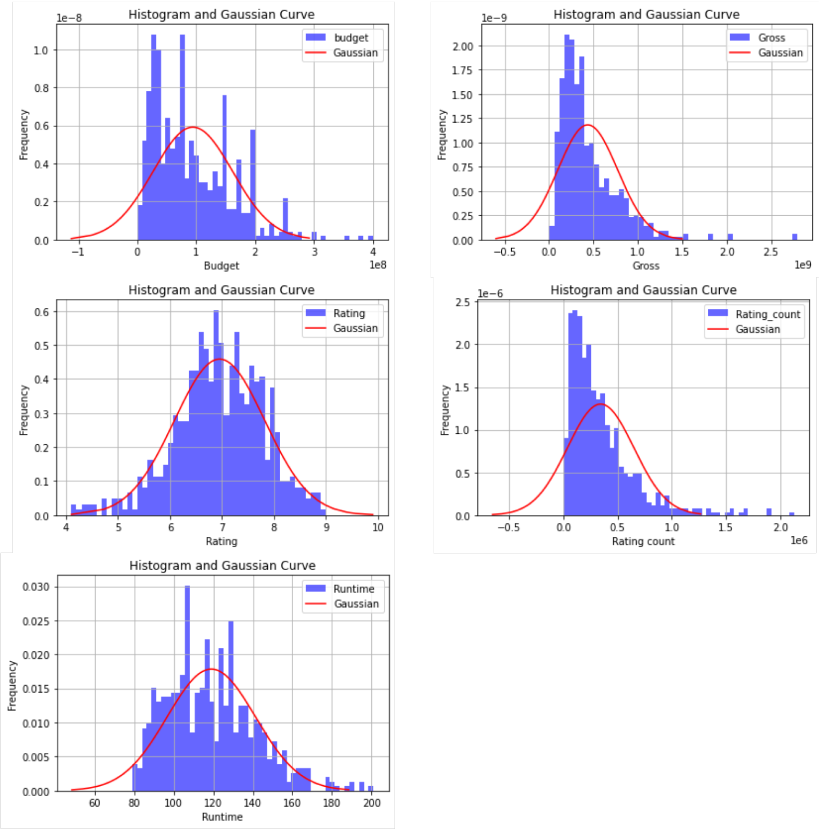

Figure 2: Frequency histograms of the features

As visible in Figure 2, only the rating feature appears to be pseudo-normally distributed, whereas *budget, gross, runtime, rating* and *rating count* have a more random distribution.

3.3 Correlation between the variables

Concerning the correlation between the variables in our dataset, we initially generated a scatterplot matrix (figure 9 in the Appendix), which is useful for visualising the relationships between features in a dataset, showing the distributions of individual features on the main diagonal and the scatter plots between pairs of features in the cells outside the diagonal. It is already visible to the naked eye that the two most correlated features are the gross-budget and rating-rating_count pairs, but the release_date-budget pair also shows some form of correlation. This correlation is confirmed by the values of the correlation matrix, where each cell contains the correlation coefficient between two features. This coefficient represents a measure of the degree of linear correlation between two variables. It takes values between -1 and 1, where these extremes represent a perfect linear correlation and it is calculated as

$$\mathrm{cor}(x, y) = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y}$$

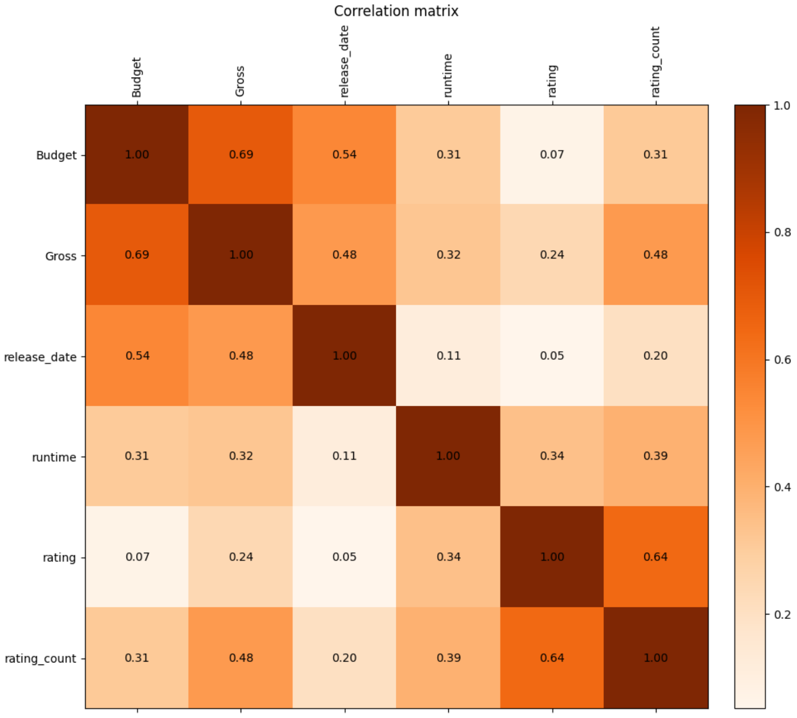

Figure 3: Correlation Matrix

As can be seen in Figure 3, the correlation coefficients with the greatest value, excluding the values in the left diagonal that indicate the correlation of a feature with itself, are precisely given by the pairs:

- $cor(budget, gross) = 0.69$
- $cor(rating, rating\_count) = 0.64$
- $cor(release\_date, budget) = 0.54$

This suggests that there is a generally positive relationship between the budget invested in film production and box office earnings. It also indicates that films well-received by critics or the audience can benefit from greater visibility and audience engagement. Furthermore, the correlation between release date and budget suggests that over time, there has been a trend towards increasing budgets in the film industry. This may reflect the evolution of the film industry, with increasingly costly and ambitious productions.

3.4 PCA

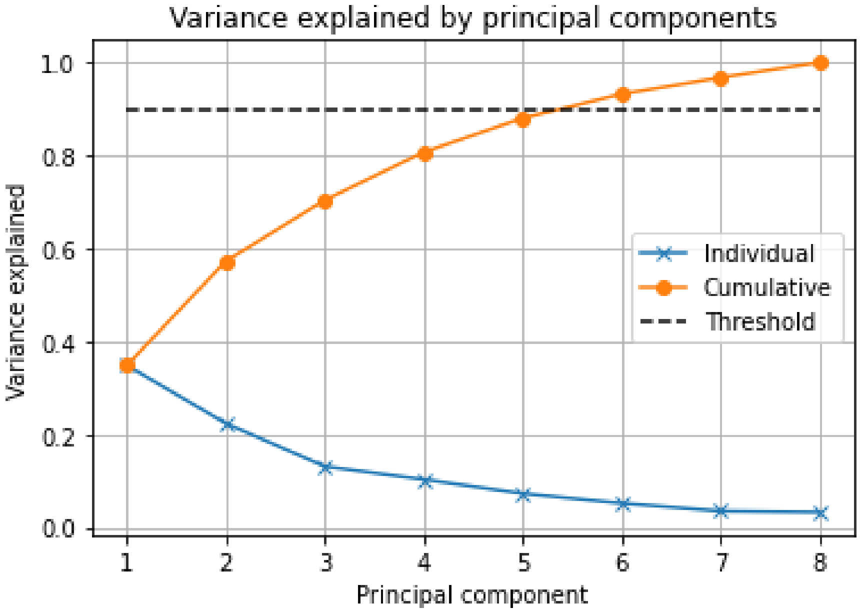

Figure 4: Variance Explained by the principal components

Before carrying out PCA on the features in our dataset, a standardisation was carried out to eliminate dependence on their scale. The standard deviation for each feature is visible in section 2.3. Once the PCA was performed, we plotted the graph showing the variance explained by the principal components (Figure 4). As derivable from the figure, the first 4 Principal Components explain around 80% of the variance.

- PC1 explains slightly below 40% of the variance and it seems to be a balanced combination of various original features without any particular feature standing out significantly (Figure 5). This suggests that the variance captured by this principal component is evenly distributed among the variables. Due to the balanced nature of the principal component, we are not able to attribute a specific meaning to this principal component in terms of the original features.

- PC2 explains slightly above 20% of the variance and it mainly consists of Gross (negative) and MPAA_rating (positive), but also negatively takes Budget into account. This component may reflect a relationship in which movies with lower budgets and earnings (indicated by the negative values of budget and gross) tend to target a more mature audience. In fact, the label encoding technique used for the MPAA rating assigns lower values to films intended for family audiences and higher values to those intended for exclusively adult audiences. This is not surprising since the horror genre has always been the preferred choice of first-time movie directors who want to make a film with a very low budget.

- PC3 explains lower than 20% of the variance and the significant positive influence of the *rating_ count* feature stands out compared to the other principal components, but like in PC1 we are not able to give a specific meaning to this principal component in terms of the original features.

- PC4 explains lower than 20% of the variance and just like PC1, no individual feature prominently stands out in PC4, indicating that the variance captured by this principal component is evenly distributed across the variables. As with PC1, PC4's balanced nature makes it challenging to attribute a specific meaning to this principal component in terms of the original features. However, the *rating* feature has a particularly high negative relevance within PC4 compared to the same feature in the other principal components.

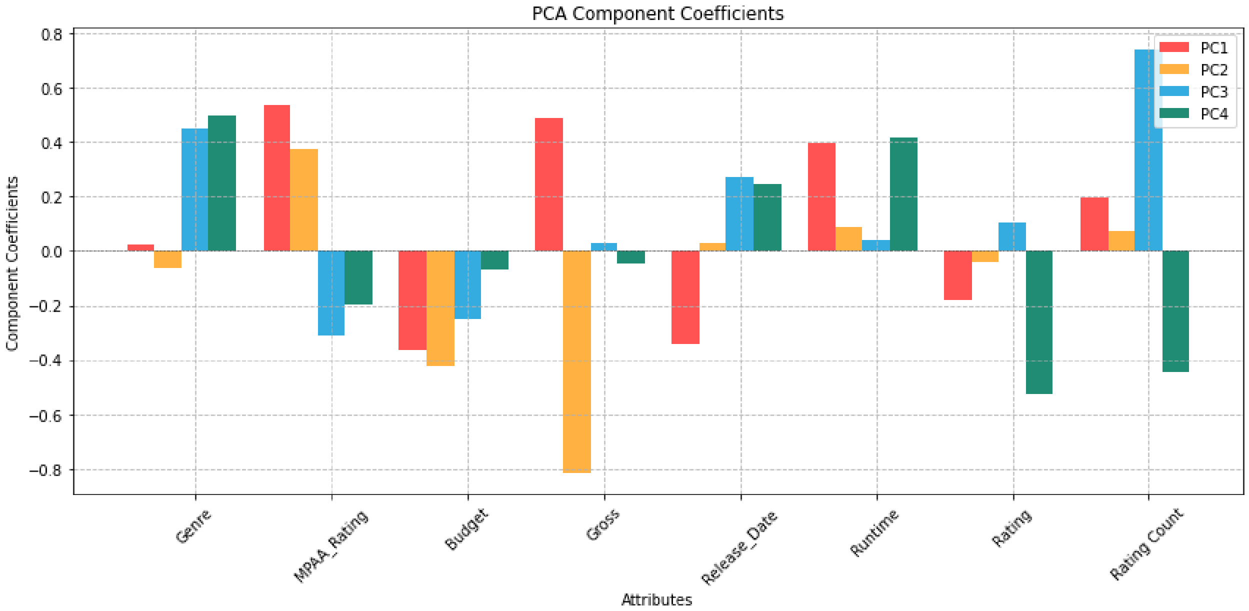

Figure 5: Representation of the coordinates of the principal components

Once PCA was performed, the entries in our dataset were projected onto the selected principal components. To enable a graphical visualization of the data, the movies were projected onto the first 2 principal components, which collectively explain approximately 60% of the variance.
As visible in Figure 6, in the right plot the points have been coloured based on their genre, while in the left one, they have been coloured according to the MPAA_rating assigned to the movies.
The various genre and MPAA_Rating classes do not separate optimally along the first 2 principal components, making it challenging to immediately interpret the data. Therefore, we decided to attempt to reduce the number of genres, considering only Action, Comedy, Horror, and War. Similarly, for MPAA_Rating, we tried to focus only on PG (Parental Guidance Suggested) and R (Restricted). The result is visible in Figure 7, and it is evident how the separation of classes is now more distinct and suitable for future classification purposes.

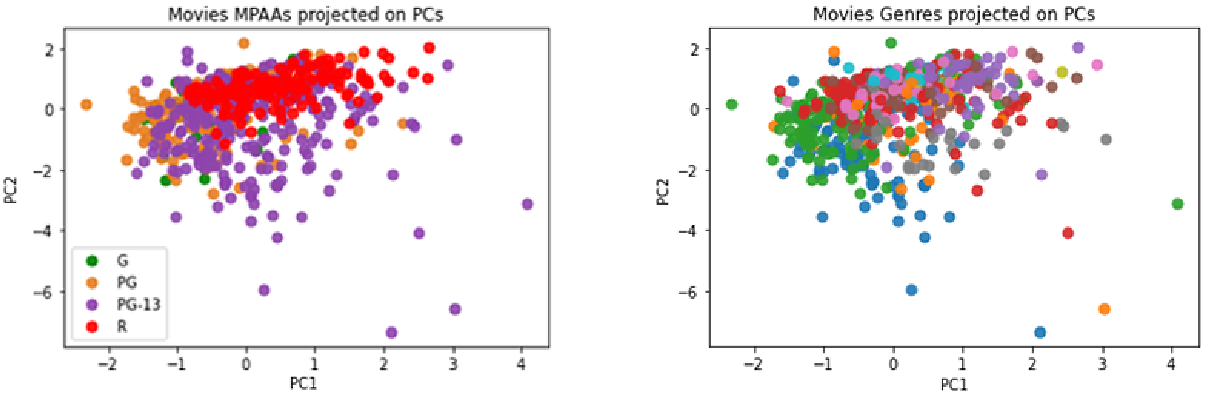

Figure 6: Movies projected on the first 2 PCs, with MPAA_Rating coloured on the left, and Genre coloured on the right

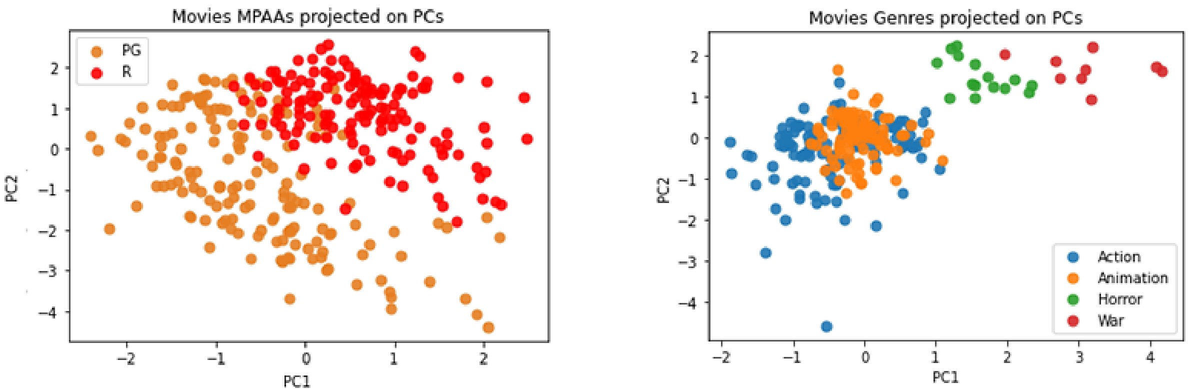

Figure 7: Movies projected on the first 2 PCs, taking into account only 4 genres and 2 MPAA Ratings

4 A discussion explaining what you have learned about the data

The dataset shows considerable variability among its attributes. *Budget* and *gross* earnings have wide-ranging figures, indicating diverse financial outcomes in the movie industry. In contrast, attributes like *runtime* and *rating* present more consistent patterns. When IQR and standard deviation are close, as seen in *runtime* and *rating*, it suggests a relatively balanced distribution around the mean. This indicates that the data for these attributes may exhibit a more symmetrical or normal distribution. However, in the case of *budget* and *gross*, high IQR and standard deviation point to significant variation and potential outliers. Despite their anomalous values, we have chosen to include these outliers as valid data points for our analysis because we think they reflect to some extent significant trends and important changes in the film industry.

Regarding the correlation between features, some interesting patterns have emerged. Firstly, the high correlation between budget and gross, with a coefficient of 0.69, unsurprisingly suggests that high-budget films tend to generate higher box office earnings, likely due to significant payments to popular actors, global marketing campaigns, and investments in cutting-edge visual and special effects. However, a higher budget does not guarantee better ratings (correlation coefficient = 0.07), emphasizing the importance of a compelling story or concept over costly facilities and technologies. The other two highly correlated feature pairs are *rating* - *rating_ count* and *release_ date* - *budget*. The first pair suggests the tendency for films with high ratings to generate more interest, positive word-of-mouth, and promotion, leading to more ratings. The second pair, on the other hand, unsurprisingly shows that more recent films have higher budgets compared to those from earlier times. Furthermore it's also interesting to notice that the lack of correlation between *release_ date* and *rating* (correlation coefficient = 0.05) implies that older films can compete with newer ones in terms of audience appeal.

Considering the regression goals specified at the beginning of the report, if we look at the correlation matrix we can generally hope that regression applied to the feature *gross* to predict a film's box office earnings based on the other features is feasible, given the interesting values of correlation coefficients between *gross* and each of the other features.

Predicting a film's rating based on all the other features, on the other hand, appears to be more challenging, considering the low correlation coefficients within the matrix.

Moving on to PCA, the heart of the project, the dataset reveals that the first four principal components collectively explain approximately 80% of the variance. We then projected the dataset entries along the first 2 principal components (that explain about 60% of the variance), highlighting the *genre* and *MPAA_ Rating* of each film. The observations did not appear to separate optimally along the first 2 principal components, making it challenging to immediately interpret the data. Therefore, we decided to reduce the number of genres, taking into consideration only Action, Comedy, Horror and War, and similarly for the *MPAA_ Rating* we tried to focus only on PG (Parental Guidance Suggested) and R (Restricted). With this restriction, the data seemed to cluster more visibly. Therefore, we believe that our initially planned classification objectives may only be partially achievable by reducing the number of classes for classification.

Overall we can say that the correlations and trends identified suggest that there are interesting relationships among key variables such as *budget*, *gross* and *rating.* However, the principal component analysis has revealed some challenges in separating the classes, which may require additional data preprocessing steps to achieve optimal results in future predictions or the reformulation of some of our classification and regression goals.

5 Exam problems for the project

1. **C**, because Time of the day is an ordinal type of data, thus this excludes all the other options.
2. **A**, because given the p-norm distance formula, with $p = \infty$,

   $$d_\infty(\mathbf{x}, \mathbf{y}) = \max_i(|x_i - y_i|)$$

   the p-norm distance between the 2 vectors is 7.
3. **A**, because using the explained variance formula

   $$explained - variance = \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{j=1}^{m} \sigma_j^2}$$

   the variance explained by the first 4 principal components is greater than 0.8
4. **D**, because looking at the principal component column specified in the answer, the option D is the only one in which high values (*Broken Trucks*, *Accident Victims* and *Defects*) correspond to positive values in the respective rows. Thus, the projection on the principal component n.2 will be a positive value.
5. **A**, because given the Jaccard similarity formula, the result is:

   $$J(s1, s2) = \frac{|s1 \cap s2|}{|s1 \cup s2|} = \frac{2}{13} = 0,153846$$

6. **B**, because

   $$p(\hat{x}_2 = 0|y = 2) = p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) + p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2) = 0.81 + 0.03 = 0.84$$

   The values come from the provided table.

   We also tried an another approach to solve the problem, however it did not work out.
   1. step: Applying the Bayes theorem

   $$p(\hat{x}_2 = 0|y = 2) = \frac{p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) * p(\hat{x}_2 = 0)}{p(y = 2)}$$

   2.step: Applying first sum rule, then marginalization

   $$p(y = 2|\hat{x}_2 = 0) = p(y = 2|\hat{x}_2 = 0, \hat{x}_7 = 0) + p(y = 2|\hat{x}_2 = 0, \hat{x}_7 = 1)$$

   $$\frac{p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) * p(y = 2)}{\sum_{i=1}^{4} p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = i) * p(y = i)} + \frac{p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2) * p(y = 2)}{\sum_{i=1}^{4} p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = i) * p(y = i)} = 0.359$$

   3.step: We should determine p(x2=0) value but we could not manage to do it

A Appendix

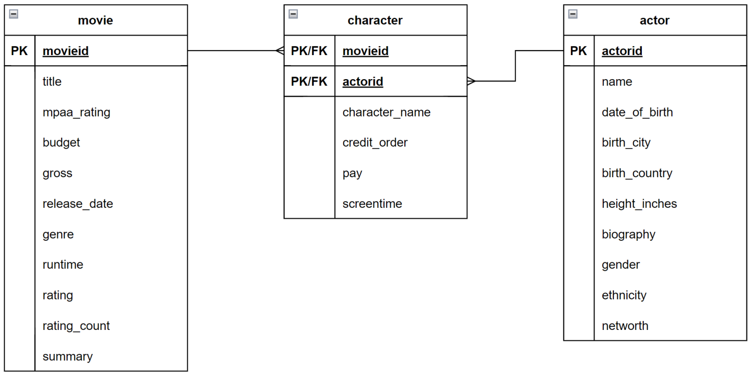

Figure 8: The original dataset

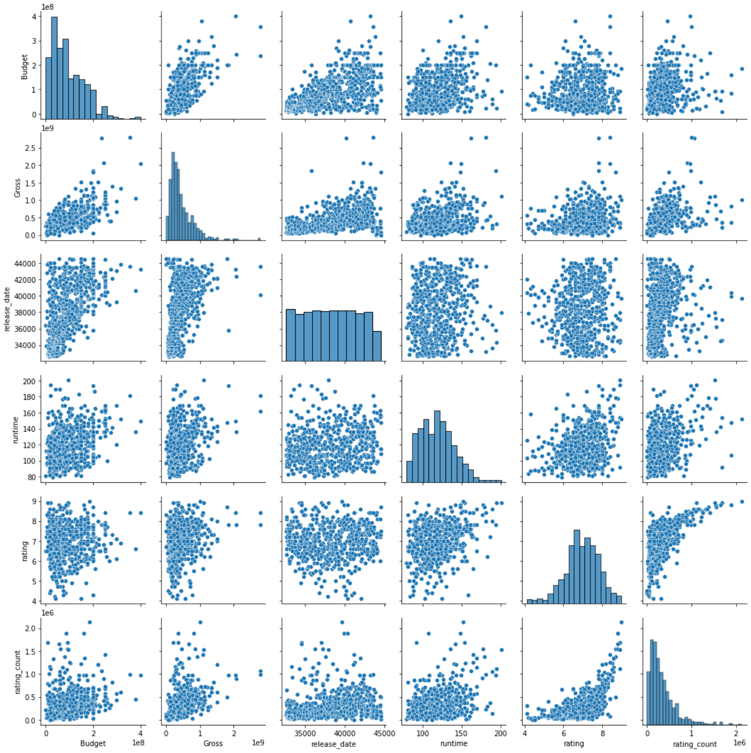

Figure 9: scatterplot matrix