Introduction to machine learning
Report I: feature extraction and visualization

Kristin Anett Remmelgas s203129
Kristo Karl Aedma s205738
Andreas Lau Hansen s194235

September 2020

Contents

1 Description of the data set

1.1 Overall problem of interest

The project will analyse the IMDb movie data set [1]. The original data set has been scraped from IMDb (the Internet Movie Database) website [2] on 01/01/2020. It consists of 85 855 movies with more than 100 votes on the website. Every object has 22 attributes. The attributes include for example the title, year of release, duration, average vote etc. A more detailed overview of the attributes can be seen in Table 1. The goal of the project is to analyse how different attributes are related to each other and how they affect the rating of the movie.

1.2 Summary of previous data analysis

Previously this data set has been used for analysing Netflix movies [3], creating a recommendation system [4], describing pandas package functionality [5] and showing some basic statistics with matplotlib [6].

The highest rated notebook on Kaggle [3] that uses this IMDb movie data set joins it with a Netflix movie data set with the goal to find movies that are both available on Netflix and have ratings on IMDb. There are some duplicates in the data according to title, release year and rating which are removed. They then find the top rated 10 movies on Netflix and top 10 movie content creating countries. Most of the further analysis is based on Netflix series and does not concern the data set used in this project.

Another notebook [4] uses the IMDb data to create a recommendation system based on content. As part of data preprocessing all missing values are imputed, relevant columns are extracted, all columns are converted into lower case and all names are split into comma separated versions. Unlike the goal of the current project this analysis focuses more on text based attributes (creating a bag of words). The end program lets the user insert a movie title and uses cosine similarity to recommend similar movies.

In conclusion the data set has previously been used for various purposes but most of the notebooks visualise the data without a more detailed analysis. A notebook that implements classification or regression on the data could not be found. The previous analysis give a good overview of data issues (missing values, duplicates, different variable types) that need to be addressed.

1.3 Machine learning tasks

Classification
For the classification task the goal is to predict a movie genre using other attributes. The difficulty with this approach is that in the original data set each movie can have multiple genres. As a solution it is possible to assign every movie only one (for example the first) genre but as the genres are listed in alphabetical order this is not the best solution. Another way is to assign every movie a random genre from the multiple that are listed. Alternatively it is possible to choose a genre that is well represented (for example drama) and try to predict whether a movie belongs in that genre or not. In order to perform classification on the data set the genre attribute should be converted from a list of string values to a one out of K encoding representation. So there is a 1 if a movie belongs to that genre and a 0 if it doesn't.

Regression
The goal for the regression task is to predict the rating (average vote) of a movie using other attributes. Initially we are planning on not including the metascore rating in this analysis because it is quite highly correlated with the average vote (as will be shown in section 3.3) and it does not make much sense to predict one rating from another. Later we will consider including the metascore as well and analysing how much the outcome improves.

2 Data attributes

2.1 Description of the attributes

The attributes in the original data set can be seen in Table 1. The table includes a short description of the attribute, it's datatype, the percentage of missing values and also the number of unique values.

For the purpose of this project some attributes where removed from the original data set that were not necessary in the context of the goal of this project. In order to keep the analysis more focused it was

decided to exclude all text based attributes except genre. That includes title, original title, country, language, director, writer, production company, actors and description. Some other attributes were excluded due to data issues described below.

Table 1: Attribute description

Attribute	Description	D. type	% of null values	No. of unique values
imdb_title_id	A unique ID for each movie	object	0.00	85855
title	Title in Italian (if possible)	object	0.00	82094
original_title	–	object	0.00	80852
year	Release year (int & string)	object	0.00	168
date_published	Release date in Italy	object	0.00	22012
genre	Genre (can be multiple)	object	0.00	1257
duration	Duration in minutes	int64	0.00	266
country	Country/ies of origin	object	0.07	4907
language	–	object	0.97	4377
director	Name of director/s	object	0.10	34733
writer	Name of writer/s	object	1.83	66859
production_company	Name of production company	object	5.19	32050
actors	List of actor names	object	0.08	85729
description	–	object	2.46	83611
avg_vote	Average vote on IMDb	float64	0.00	89
votes	Number of votes on IMDb	int64	0.00	14933
budget	Budget in mixed currencies	object	72.38	4642
usa_gross_income	Gross income in USA	object	82.15	14857
worlwide_gross_income	Gross income worldwide	object	63.87	30414
metascore	Rating on metacritic.com	float64	84.50	99
reviews_from_users	Number of reviews	float64	8.85	1213
reviews_from_critics	Number of reviews	float64	13.74	595

2.2 Data issues

Spurious attributes
The IMDb movie ID attribute is irrelevant in the context of this project. It is used only as an identification for a movie and does not give any valuable information therefore the attribute was excluded from further analysis.

Missing values
In the case of some attributes over 50 % of the values were missing [5]. Such attributes are metascore, worldwide gross income, USA gross income and budget. All of these except metascore were excluded from analysis. Metascore values describe movie rating on Metacritic webpage [7]. It was decided that this is something we would definitely like to include in our analysis and also due to it's nice distribution the attribute was kept even though it had a lot of missing values. The movies which didn't have a metascore rating were therefore excluded which decreased the number of analysed movies considerably. It is not possible to impute any values, as the movies are completely unrelated to each other.

Variable type issues
Worldwide gross income, USA gross income and budget introduce the problem of having different values. Some values are numeric but due to the fact that different currencies are used some contain the numeric value and a character representing the currency [5].

Release year attribute has some issues when reading in the data set. Most of the values are classified as integer values (as it should be) but about third of the values are represented as string values even though only one of them is actually a string ("2009 TV Movie"). After fixing this year value it was possible to make all year values into integers.

The Italian heritage of the data set
As the data set scraping has most likely been done in Italy it has left a mark on the data set. Attribute

title has several titles in Italian even though the `original_title` might be in another language and Italy is not listed among the origin countries.

This also has an effect on the `date_published` attribute. On several occasions it can be seen that it doesn't match with release year and on some cases the difference is several years. After further investigation on IMDb webpage it was discovered that the `date_published` value quite often matches the date when the movie was released in Italy. Sometimes it is a random country's release date (for example when there wasn't Italy's release date recorded).

Duplicates
In previous analysis the data set has been cleaned of some duplicates based on title, release year and IMDb score [3] but in the final data set (after excluding unnecessary attributes) none were found.

2.3 Data set after preprocessing

As previously described there are a lot of attributes we don't need or have issues with therefore many attributes were excluded from further analysis. As it is a large data set and there is no shortage of movies any object (movie) that had missing values on the attributes that we were interested was dropped too. The final attributes and their types can be seen in Table 2.

Attributes kept
We mainly kept data that already had numerical (int64 and float64) values. Such values are duration, avg_vote, votes, metascore, reviews_from_users and reviews_from_critics. This includes year values after they were fixed.

Secondly, we kept the attribute for movie genres. It had the genres for the movies as strings. If a movie had more than one genre it was all one string and the genres were separated with comma and whitespace (", "). Knowing that, we were able to split the strings and turn the whole genres attribute into 1-out-of-K encoding. Now the added attributes are names of genres and have values 1 or 0 whether movie had such genre listed or not. 1-out-of-K encoding for genres was done after dropping unsuitable columns and rows.

After processing, the data set consists of 9294 movies with 27 attributes out of which 20 are genres using 1-out-of-K encoding.

Table 2: Attributes after preprocessing

Attribute	Discrete or continuous	Type
Year	Discrete	Interval
Duration	Continuous	Ratio
Average vote	Continuous	Ordinal
No of votes	Discrete	Ratio
Metascore	Continuous	Ordinal
No. of reviews from users	Discrete	Ratio
No. of reviews from critics	Discrete	Ratio
Genre	Discrete	Nominal

2.4 Basic summary statistics of the attributes

Table 3 shows some basic summary statistics of each attribute. Note that for the reviews and votes there is a rather large difference between the mean and the median, as there are quite a few movies with over a million votes.

The data set is not perfectly balanced in terms of the size of each label of the movie. Figure 1 shows the distribution of the genres. It is clear that drama and comedy are over represented in the data set, while both westerns and musicals are almost non-existent.

Table 3: Summary statistics of the attributes

	Year	Duration	Avg. vote	Votes	Metascore	Users r.	Critics r.
Mean	2011.3 1	103.81	6.23	51,077.31	54.73	191.01	113.91
Std.	5.38	19.82	0.95	118,784.7	17.18	407.33	118.92
Min	2001	50	1.4	103	1	1	1
Median	2012	100	6.3	8665	55	60	71
Max	2020	808	9	2,241,615	100	10472	999

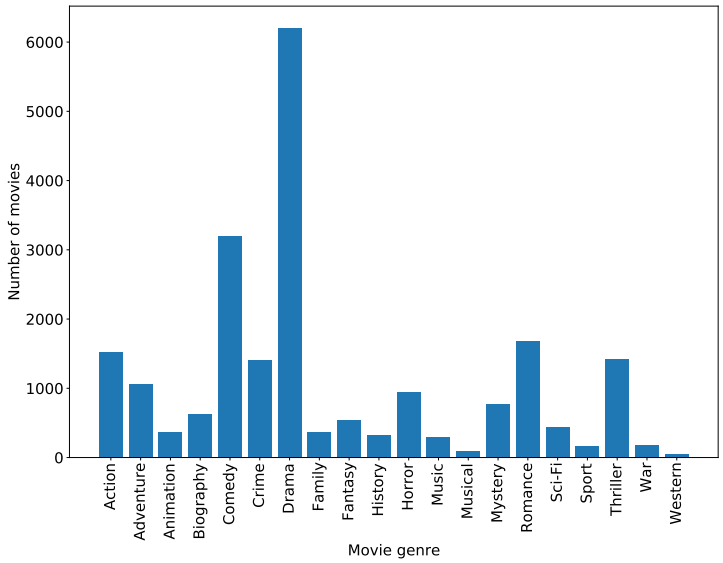

Figure 1: Frequency of different genres

3 Data visualisation

3.1 Outliers

The data set contains some values that differ greatly from others of the same attribute. One such example is a movie with a duration of 808 minutes (13.5 hours). As it is a real movie and a correct duration this was not classified as an outlier. Figure 2 shows histograms for all the attributes, especially the histograms for the duration and votes as well as the number of reviews from users indicate that there might be outliers. While the histograms attenuate quite quickly there are still a significant number of movies in the flat areas. None of them are outliers per say, as they follow from each other.

3.2 Distributions

We want to know how the different attributes are spread out according to their values in order to find out if there is a systematic pattern or any concerning trends in the data. Figure 2 shows this. It seems like we have more movies in the more recent years, with 2020 being half all the other years, due to the data set collection date. For the votes and reviews a lot of movies have very few entries while some have a lot, which makes the histogram a bit compressed.

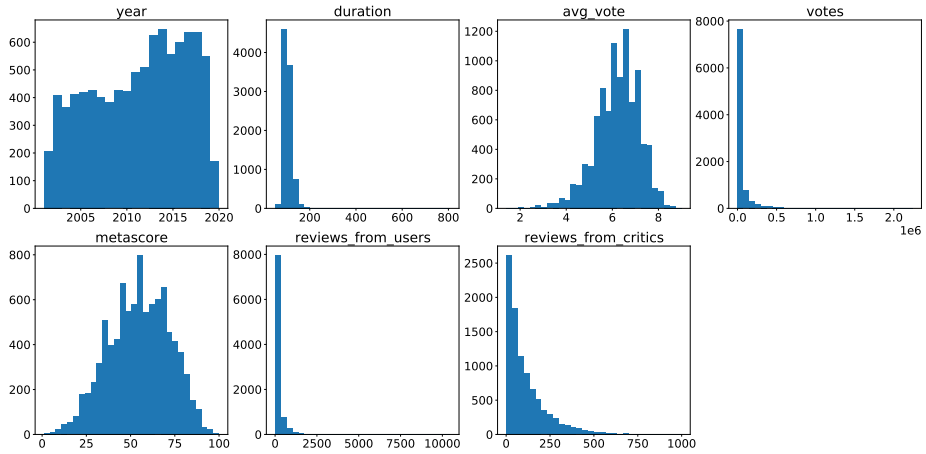

Figure 2: Attribute histograms

To investigate whether or not the attributes are normal distributed, we generated qq-plots for the duration, metascore and average votes as they are the only three attributes which are somewhat normal distributed according to Figure 2.

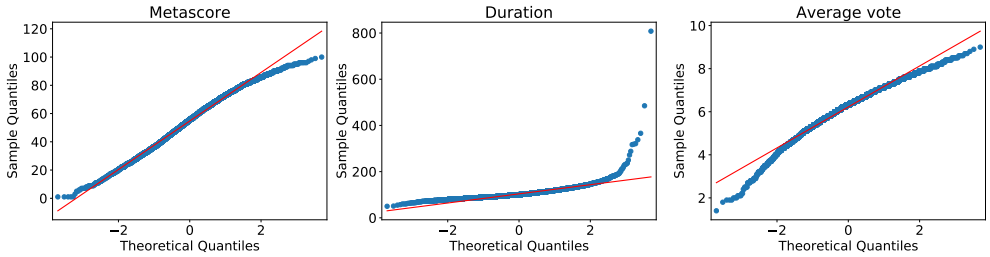

Figure 3: qq-plots for Metascore, duration and average vote. Attributes drift off from a normal distribution

The qq-plots show that the metascore and duration are almost normal distributed, as they only deviate towards the higher end of the quantiles. The average vote is way far off being normal distributed, since the curve is way too steep so only the middle values follow a normal curve and the sample quantiles are below the expected at either end.

3.3 Correlation between variables

The matrix in Figure 4 indicates that certain variables and movie genres are correlated with each other. The number of votes is highly correlated with both types of reviews this is likely because when a movie is popular more people will go to IMDb and give a vote and the reviewers are also attracted. Average vote and the metascore are unsurprisingly correlated as well as could also be seen in previous analysis of the data[5]. This makes sense given metascore is a score between 0 and 100, which is pretty much just 10x the average vote on IMDb. There are also some other interesting things like how the romance genre is negatively correlated with the year which must mean that fewer new movies are romantic than before. Then we can also see which genres are typically used in conjunction with each other like action, sci-fi, and adventure. On the other hand thrillers and comedies are not a good pairing.

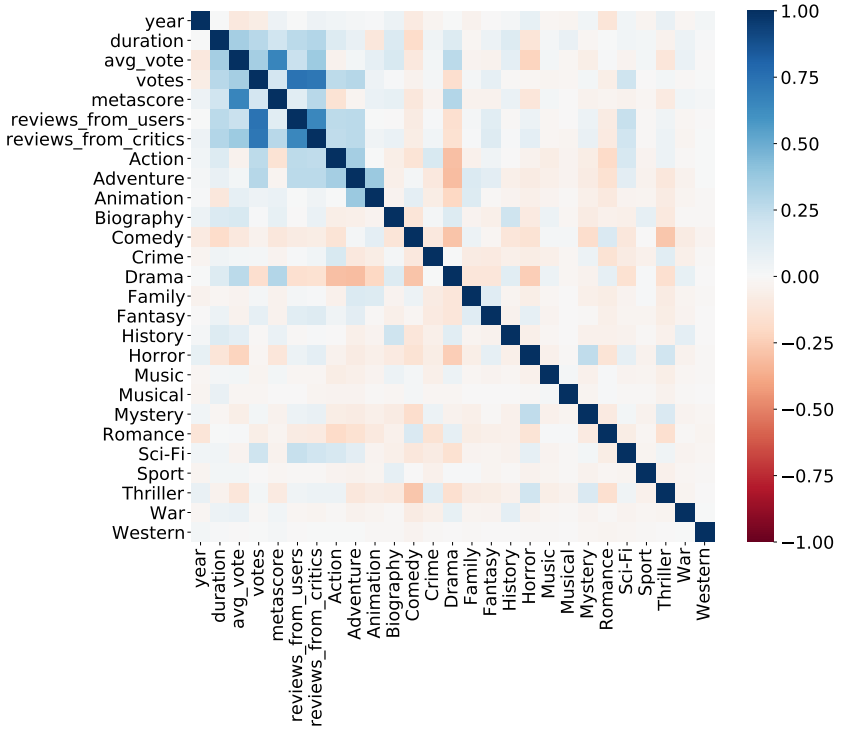

Figure 4: Correlation matrix

3.4 Principal Component Analysis (PCA)

After we performed the one-out-K encoding we were left with a data set with a dimensionality of 27. With 2- or 3d data it is easy to make scatter plots, this is not the case with our 27 dimensions. Therefore, we used PCA via singular value decomposition to reduce the dimensions to 2. Before conducting the PCA analysis the differences in scale were corrected by standardization. The standardization was carried out by ensuring each attribute has a mean of zero and unit variance.

Variation explained

The number of principal components (PC's) required to explain 90% variance of the data is 21. With the 2 we are using to plot, we can account for roughly 20% variance. Figure 5 shows the cumulative variance as a function of the number of PC's (orange) as well as the variance explained by each subsequent PC.

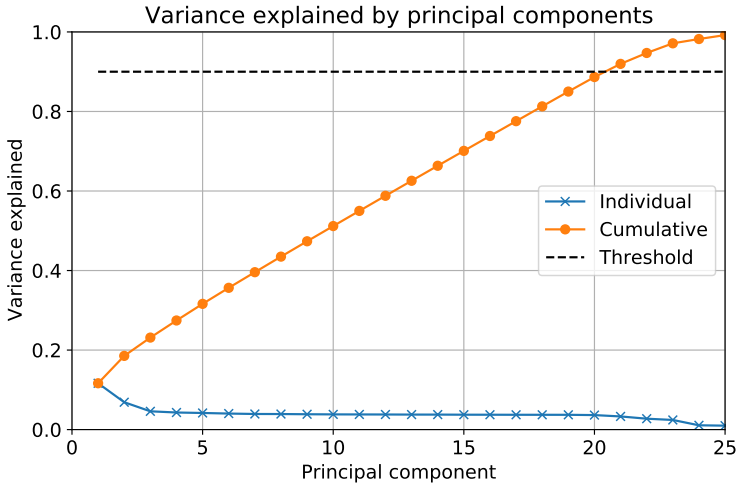

Figure 5: PCA components values separately and cumulatively, threshold is set to 0.9

Principal directions of the considered PCA components

We can interpret the principal directions of the considered PCA components by investigating the coefficients in the vectors of $\mathbf{V}$. It can be seen in Figure 6 that the first principal component mainly captures the variation of the number of votes and reviews. The best described genres are action and adventure. All these attributes also have a negative coefficient. The second principal component mainly captures the average vote, metascore and drama genre. The third principal component represents the variation of adventure, animation and comedy with a positive coefficient and horror, thriller and history with a negative coefficient. This knowledge can now be taken into account when deciding which principal components to project the data onto.

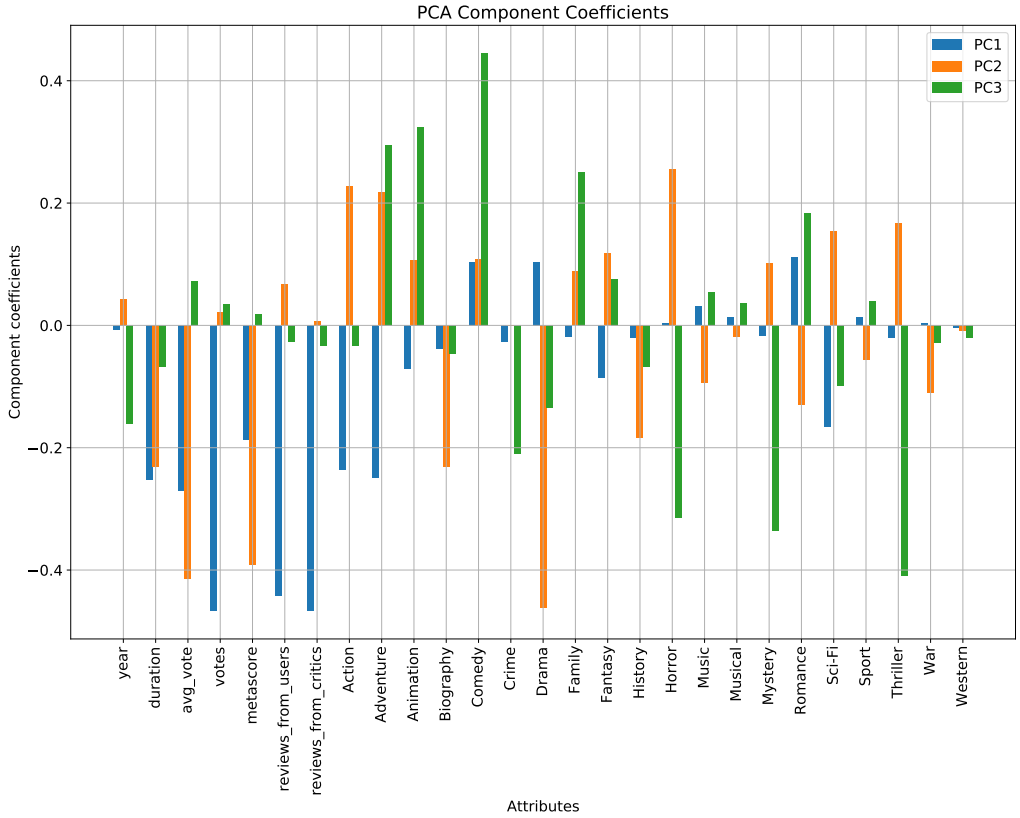

Figure 6: Barplot of PCA's first three components

Projected data

Finally we can project the data onto the principal components in order to visualise any structures and groupings in the data. In Figure 7, we have coloured the movies according to their genre. A problem here is that some of movies have several genres, in that case we chose to colour the movie according to one of the genres chosen at random.

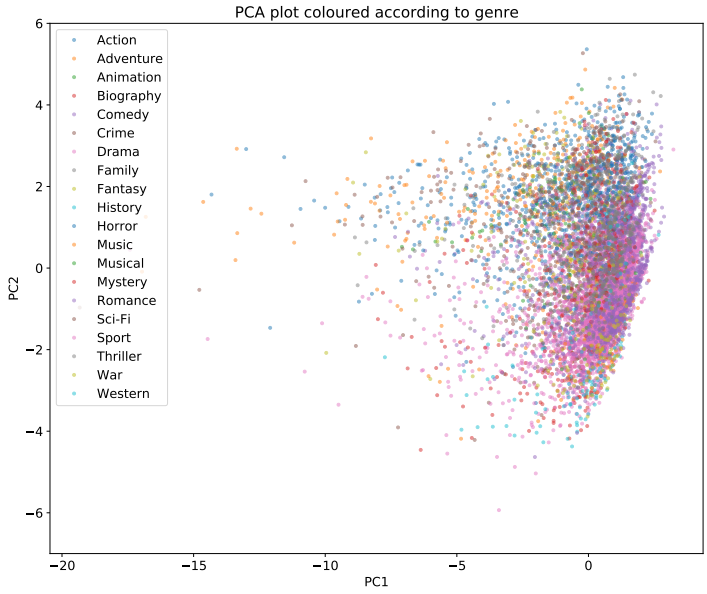

Figure 7: PCA projection on components 1 and 2

Figure 7 shows that especially the drama genre (pink) is quite well separated from the others. It is a bit of a problem that there are so many genres so it is almost impossible to discern any individual genre in the middle. Most of the genres do seem to have a slightly right tilted vertical scatter placed right next to each other. You can distinguish the romance genre (purple) as lump right next to drama.

To more clearly show overlap between the genres Figure 8 shows the same projection, but where the movies have been coloured according to whether or not they belong to the drama class. The overlap between drama and the big blob becomes more apparent with this visualisation.

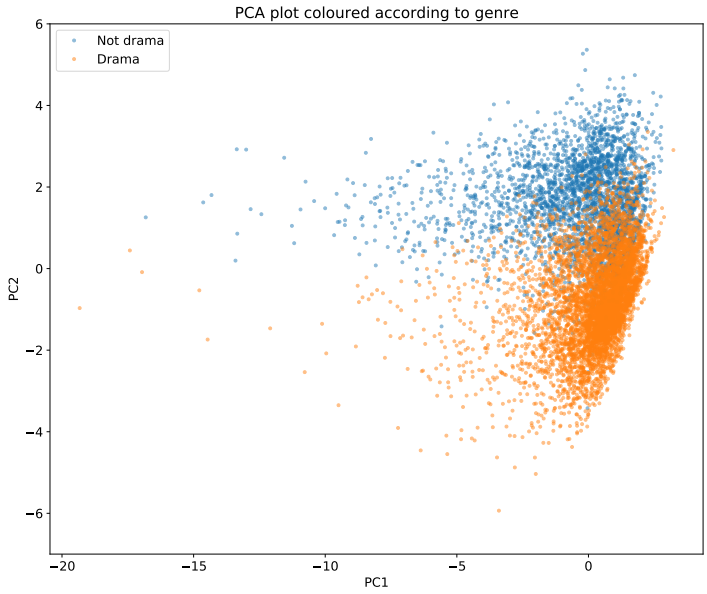

Figure 8: PCA projection with separation between movies with and without drama genre

4 Discussion

We learned about our dataset that there were a lot of missing values for some attributes what is unfortunate and that there are some string value attributes that we decided not to include in the analysing part. We learned that there are some positive correlations between attributes like metascore and average vote and some negative correlations between genres like drama and action. This is great as it might help predict our goals better. We learned that many of the chosen attributes have highly skewed distributions that have some extreme values. This means we have quite wide variance for those attributes but the majority is compressed tightly somewhere else. This was seen also really well on the PCA graphs. Analysing the principal components we learned that they all have quite similar variances and there can't be seen specific separation between data points when projecting them on the first ones. This means that we will probably have difficulties when trying to predict most genres based on other attributes.

Overall it seems that predicting metascore values as a regression problem seems highly doable when including IMDb's average vote but might be difficult to do it when excluding avg_vote from inputs. For classification problem it seems that predicting a movie's genre might be a bit troublesome for some genres considering how the PCA plots are looking. In the PCA plots it is difficult to make any obvious decision boundaries, however this might not be a problem considering the fact that when we are going to do the classification we will have access to all PCA dimensions.

5 Contributions

Table 4: Contributions

Section	**40% or more responsible**
1.1 Overall problem of interest	Kristin
1.2 Summary of previous data analysis	Kristin
1.3 Machine learning tasks	Kristin
2.1 Description of the attributes	Kristin
2.2 Data issues	Kristin, Kristo
2.3 Dataset after preprocessing	Andreas, Kristo
2.4 Basic summary statistics of the attributes	Andreas, Kristo
3.1 Outliers	Kristin
3.2 Distributions	Andreas
3.3 Correlations	Andreas
3.4 PCA analysis	Andreas
4. Discussion	Kristo

References

[1] URL: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset.

[2] URL: https://www.imdb.com.

[3] URL: https://www.kaggle.com/niharika41298/netflix-visualizations-recommendation-eda.

[4] URL: https://www.kaggle.com/sasha18/recommender-systems-based-on-content.

[5] URL: https://www.kaggle.com/goldens/imdb-pandas-review.

[6] URL: https://www.kaggle.com/stefanoleone992/imdb-eda.

[7] URL: https://www.metacritic.com/.