POZNAN UNIVERSITY OF TECHNOLOGY

# Movies’ genres, titles and ratings association rules mining.

Mikołaj Marmurowicz (Mickeyo0o), Kajetan Sulwiński (ekohachi22)

2023-06-12

## Contents

# Executive Summary

In this report we will analyze the association rules that can be derived from movies and their ratings. Specifically, we will test:

- are there associations between movie genres,
- are there specific words in titles that can be associated with genres,
- are movies‘ ratings in any way associated with their genres,
- are there any association of genres liked by users.

Our experiments have proven that for all above assumptions association rules can be found.

The first experiment shows that association rules can be found where expected, e.g.
**Mystery and Horror -> Thriller**.
The second experiment depicts slightly less obvious information, that there are some specific words in titles that can be directly associated with genres, e.g.
**love -> Romance**.
The third experiment manifests important information regarding the film industry, where we can notice that specific genres of movies can be associated with an average rating above or below a given value, e.g.
**Action -> Average rating worse than 3.5**.
The last experiment reveals that there exist association rules between genres that specific users like together, e.g.
**Adventure -> Action**.

We conclude that genres of movies that are similar in nature can be easily inferred from association rules, these rules are very closely related to the rules that present genres that are liked together. That means that people usually tend to like the same or only slightly different genres of movies, and do not rate movies outside of their usual choice.
What is more, average rating of a movie inferred from association rules usually suggests that a given genre will receive a mark worse than 4.0, that is due to the fact that only the most popular genres were presented in the association rules, thus making the actually good films even more rare. That is because the more popular the genre is the more films are produced with given genre, and with quantity sadly the quality usually does not follow. That means that most films produced in popular genres are not really good and only a few big productions receive good ratings, meaning that the support of such grades is very low.
Additionally, we have discovered a few words that are connected to specific movie genres, most of them are logical, but there are quite a few interesting inferences, such as that Comedy and action movies are most likely to receive a second part, as the association rules show that “2” or “II” is most oftenly found in titles related to these genres.

# Introduction

This report is produced as an additional fix for the second assignment of the Data Mining course. We will try to find interesting association rules based on the movies and their ratings. Specifically, we will delve into the association between genres, whether the title words suggest some sort of genre, whether genre suggests some sort of grade, and what users usually like together.

# Methodology

To perform following experiments we have used a data set given at our laboratories that represent movies and ratings of them.

Additionally, to perform all necessary calculations we have used the following libraries in Python:

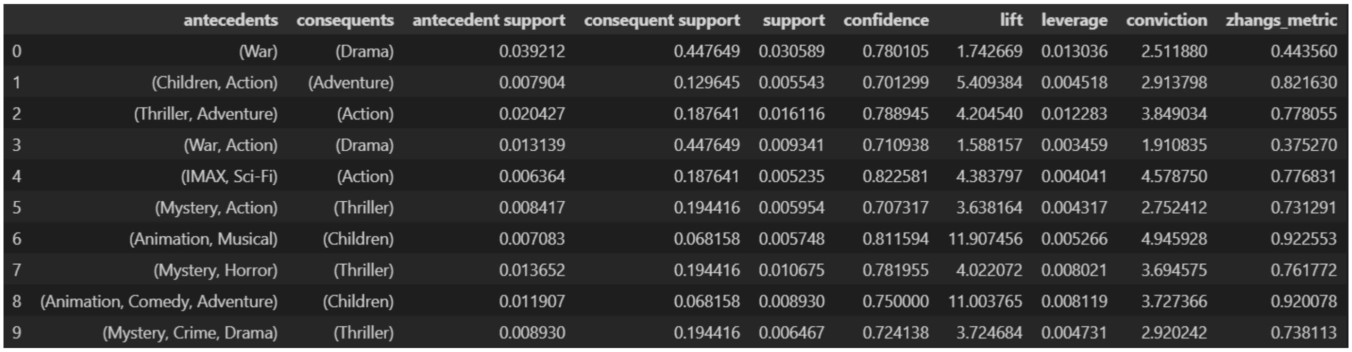
- pandas - used to create data frames and perform transformations
- numpy - used to perform more complex calculations more efficiently
- sklearn.preprocessing MultiLabelBinarizer - used to binarize the data frames
- mlxtend.frequent__patterns apriori - used to apply the apriori algorithm
- mlxtend.frequent__patterns association__rules - used to create association rules
- nltk.corpus stopwords - used to remove stopwords from titles

From the data set column timestamp is dropped as it will not be used in the experiments. What is more, for each genre a new column has been created, and each movie genres have been binarized into these columns. Additionally, English stopwords have been removed the titles of movies. Additional information has been added to the movies data frame that represents the ratings average value, and a new data frame was created to perform the last test that combines all users and the movies they liked.

# Results & findings

## Test 1

### Are there associations between movie genres?



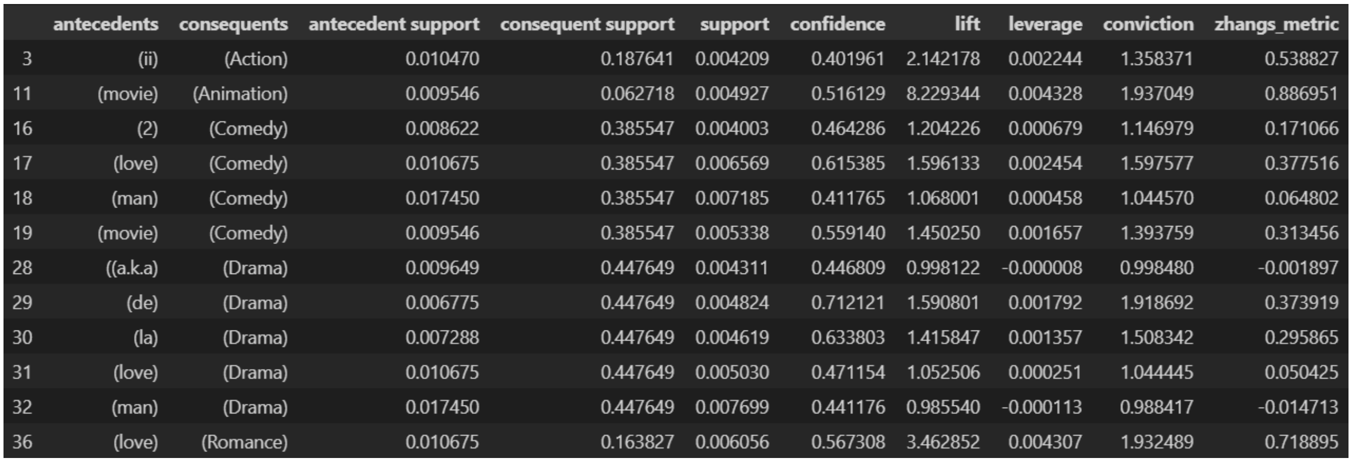
| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | (War) | (Drama) | 0.039212 | 0.447649 | 0.030589 | 0.780105 | 1.742669 | 0.013036 | 2.511880 | 0.443560 |
| 1 | (Children, Action) | (Adventure) | 0.007904 | 0.129645 | 0.005543 | 0.701299 | 5.409384 | 0.004518 | 2.913798 | 0.821630 |
| 2 | (Thriller, Adventure) | (Action) | 0.020427 | 0.187641 | 0.016116 | 0.788945 | 4.204540 | 0.012283 | 3.849034 | 0.778055 |
| 3 | (War, Action) | (Drama) | 0.013139 | 0.447649 | 0.009341 | 0.710938 | 1.588157 | 0.003459 | 1.910835 | 0.375270 |
| 4 | (IMAX, Sci-Fi) | (Action) | 0.006364 | 0.187641 | 0.005235 | 0.822581 | 4.383797 | 0.004041 | 4.578750 | 0.776831 |
| 5 | (Mystery, Action) | (Thriller) | 0.008417 | 0.194416 | 0.005954 | 0.707317 | 3.638164 | 0.004317 | 2.752412 | 0.731291 |
| 6 | (Animation, Musical) | (Children) | 0.007083 | 0.068158 | 0.005748 | 0.811594 | 11.907456 | 0.005266 | 4.945928 | 0.922553 |
| 7 | (Mystery, Horror) | (Thriller) | 0.013652 | 0.194416 | 0.010675 | 0.781955 | 4.022072 | 0.008021 | 3.694575 | 0.761772 |
| 8 | (Animation, Comedy, Adventure) | (Children) | 0.011907 | 0.068158 | 0.008930 | 0.750000 | 11.003765 | 0.008119 | 3.727366 | 0.920078 |
| 9 | (Mystery, Crime, Drama) | (Thriller) | 0.008930 | 0.194416 | 0.006467 | 0.724138 | 3.724684 | 0.004731 | 2.920242 | 0.738113 |

The following results were achieved by setting the minimum support to 0.005 and minimum confidence to 0.7.

We can see that the association rules represent movies which by humans are considered to be similar, or that are usually seen together. For example **Animation and Musical -> Children** greatly represents that most movies that are animated and contain musical elements usually are made for children. Or a rule **Mystery and Horror -> Thriller** represents the fact these rules not only show the associations, but similarity of the genres to each other.

## Test 2

**Are there specific words in titles that can be associated with genres?**



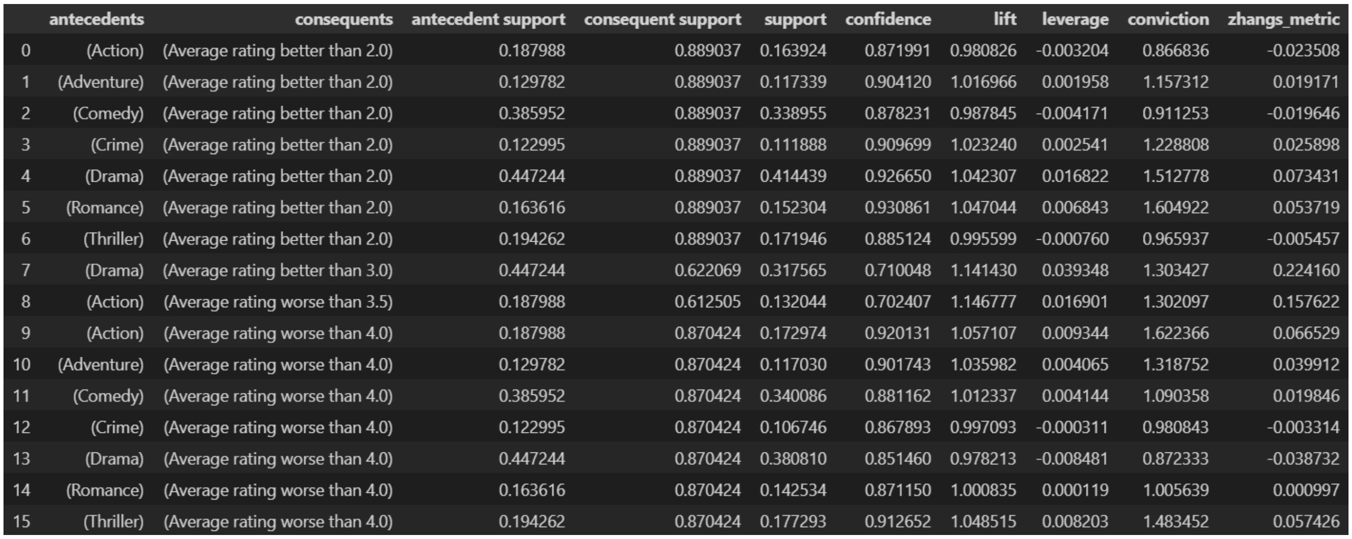
| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | (ii) | (Action) | 0.010470 | 0.187641 | 0.004209 | 0.401961 | 2.142178 | 0.002244 | 1.358371 | 0.538827 |
| 11 | (movie) | (Animation) | 0.009546 | 0.062718 | 0.004927 | 0.516129 | 8.229344 | 0.004328 | 1.937049 | 0.886951 |
| 16 | (2) | (Comedy) | 0.008622 | 0.385547 | 0.004003 | 0.464286 | 1.204226 | 0.000679 | 1.146979 | 0.171066 |
| 17 | (love) | (Comedy) | 0.010675 | 0.385547 | 0.006569 | 0.615385 | 1.596133 | 0.002454 | 1.597577 | 0.377516 |
| 18 | (man) | (Comedy) | 0.017450 | 0.385547 | 0.007185 | 0.411765 | 1.068001 | 0.000458 | 1.044570 | 0.064802 |
| 19 | (movie) | (Comedy) | 0.009546 | 0.385547 | 0.005338 | 0.559140 | 1.450250 | 0.001657 | 1.393759 | 0.313456 |
| 28 | ((a.k.a) | (Drama) | 0.009649 | 0.447649 | 0.004311 | 0.446809 | 0.998122 | -0.000008 | 0.998480 | -0.001897 |
| 29 | (de) | (Drama) | 0.006775 | 0.447649 | 0.004824 | 0.712121 | 1.590801 | 0.001792 | 1.918692 | 0.373919 |
| 30 | (la) | (Drama) | 0.007288 | 0.447649 | 0.004619 | 0.633803 | 1.415847 | 0.001357 | 1.508342 | 0.295865 |
| 31 | (love) | (Drama) | 0.010675 | 0.447649 | 0.005030 | 0.471154 | 1.052506 | 0.000251 | 1.044445 | 0.050425 |
| 32 | (man) | (Drama) | 0.017450 | 0.447649 | 0.007699 | 0.441176 | 0.985540 | -0.000113 | 0.988417 | -0.014713 |
| 36 | (love) | (Romance) | 0.010675 | 0.163827 | 0.006056 | 0.567308 | 3.462852 | 0.004307 | 1.932489 | 0.718895 |

The following results were achieved by setting the minimum support to 0.004 and minimum confidence to 0.4.

We can see that these association rules show that the words “love” or “man” are usually used in titles of movies in only specific genres such as **Drama** or **Romance**. We can also notice that only a few genres receive the same title, but with the number to represent the next part of the movie e.g. **2 -> Comedy** or **II -> Action**. Additionally, a common phenomenon in the Comedy movies industry can be noticed - we can oftenly see a word **movie** in the title, e.g. “Scary movie”.

## Test 3

### Are movies‘ ratings in any way associated with their genres?



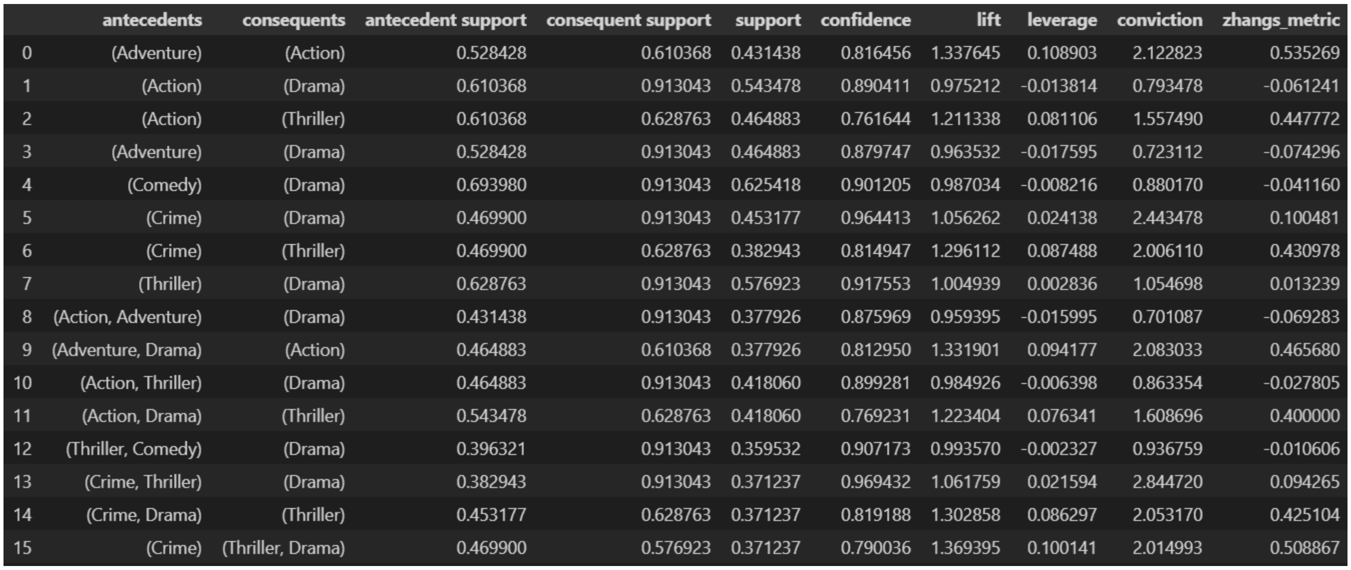
| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | (Action) | (Average rating better than 2.0) | 0.187988 | 0.889037 | 0.163924 | 0.871991 | 0.980826 | -0.003204 | 0.866836 | -0.023508 |
| 1 | (Adventure) | (Average rating better than 2.0) | 0.129782 | 0.889037 | 0.117339 | 0.904120 | 1.016966 | 0.001958 | 1.157312 | 0.019171 |
| 2 | (Comedy) | (Average rating better than 2.0) | 0.385952 | 0.889037 | 0.338955 | 0.878231 | 0.987845 | -0.004171 | 0.911253 | -0.019646 |
| 3 | (Crime) | (Average rating better than 2.0) | 0.122995 | 0.889037 | 0.111888 | 0.909699 | 1.023240 | 0.002541 | 1.228808 | 0.025898 |
| 4 | (Drama) | (Average rating better than 2.0) | 0.447244 | 0.889037 | 0.414439 | 0.926650 | 1.042307 | 0.016822 | 1.512778 | 0.073431 |
| 5 | (Romance) | (Average rating better than 2.0) | 0.163616 | 0.889037 | 0.152304 | 0.930861 | 1.047044 | 0.006843 | 1.604922 | 0.053719 |
| 6 | (Thriller) | (Average rating better than 2.0) | 0.194262 | 0.889037 | 0.171946 | 0.885124 | 0.995599 | -0.000760 | 0.965937 | -0.005457 |
| 7 | (Drama) | (Average rating better than 3.0) | 0.447244 | 0.622069 | 0.317565 | 0.710048 | 1.141430 | 0.039348 | 1.303427 | 0.224160 |
| 8 | (Action) | (Average rating worse than 3.5) | 0.187988 | 0.612505 | 0.132044 | 0.702407 | 1.146777 | 0.016901 | 1.302097 | 0.157622 |
| 9 | (Action) | (Average rating worse than 4.0) | 0.187988 | 0.870424 | 0.172974 | 0.920131 | 1.057107 | 0.009344 | 1.622366 | 0.066529 |
| 10 | (Adventure) | (Average rating worse than 4.0) | 0.129782 | 0.870424 | 0.117030 | 0.901743 | 1.035982 | 0.004065 | 1.318752 | 0.039912 |
| 11 | (Comedy) | (Average rating worse than 4.0) | 0.385952 | 0.870424 | 0.340086 | 0.881162 | 1.012337 | 0.004144 | 1.090358 | 0.019846 |
| 12 | (Crime) | (Average rating worse than 4.0) | 0.122995 | 0.870424 | 0.106746 | 0.867893 | 0.997093 | -0.000311 | 0.980843 | -0.003314 |
| 13 | (Drama) | (Average rating worse than 4.0) | 0.447244 | 0.870424 | 0.380810 | 0.851460 | 0.978213 | -0.008481 | 0.872333 | -0.038732 |
| 14 | (Romance) | (Average rating worse than 4.0) | 0.163616 | 0.870424 | 0.142534 | 0.871150 | 1.000835 | 0.000119 | 1.005639 | 0.000997 |
| 15 | (Thriller) | (Average rating worse than 4.0) | 0.194262 | 0.870424 | 0.177293 | 0.912652 | 1.048515 | 0.008203 | 1.483452 | 0.057426 |

The following results were achieved by setting the minimum support to 0.1 and minimum confidence to 0.7.

We can see that these association rules show that most movies have some sort of an average rating over the genre close to 3.0, as most popular genres are better than 2.0, but usually worse than 4.0. There are two special cases - **Drama -> Average rating better than 3.0** meaning that drama movies are associated with slightly better rating than other movies, therefore confidence was high enough. The other special case **Action -> Average rating worse than 3.5**, which shows that though it is one of the most popular movie genres, the ratings are not as good as of other genres, and the amount of data to prove that was high enough to reach given confidence. That shows that with quantity of Action movies does not necessarily come quality.

## Test 4

### Are there any associations of genres liked by users?



| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | (Adventure) | (Action) | 0.528428 | 0.610368 | 0.431438 | 0.816456 | 1.337645 | 0.108903 | 2.122823 | 0.535269 |
| 1 | (Action) | (Drama) | 0.610368 | 0.913043 | 0.543478 | 0.890411 | 0.975212 | -0.013814 | 0.793478 | -0.061241 |
| 2 | (Action) | (Thriller) | 0.610368 | 0.628763 | 0.464883 | 0.761644 | 1.211338 | 0.081106 | 1.557490 | 0.447772 |
| 3 | (Adventure) | (Drama) | 0.528428 | 0.913043 | 0.464883 | 0.879747 | 0.963532 | -0.017595 | 0.723112 | -0.074296 |
| 4 | (Comedy) | (Drama) | 0.693980 | 0.913043 | 0.625418 | 0.901205 | 0.987034 | -0.008216 | 0.880170 | -0.041160 |
| 5 | (Crime) | (Drama) | 0.469900 | 0.913043 | 0.453177 | 0.964413 | 1.056262 | 0.024138 | 2.443478 | 0.100481 |
| 6 | (Crime) | (Thriller) | 0.469900 | 0.628763 | 0.382943 | 0.814947 | 1.296112 | 0.087488 | 2.006110 | 0.430978 |
| 7 | (Thriller) | (Drama) | 0.628763 | 0.913043 | 0.576923 | 0.917553 | 1.004939 | 0.002836 | 1.054698 | 0.013239 |
| 8 | (Action, Adventure) | (Drama) | 0.431438 | 0.913043 | 0.377926 | 0.875969 | 0.959395 | -0.015995 | 0.701087 | -0.069283 |
| 9 | (Adventure, Drama) | (Action) | 0.464883 | 0.610368 | 0.377926 | 0.812950 | 1.331901 | 0.094177 | 2.083033 | 0.465680 |
| 10 | (Action, Thriller) | (Drama) | 0.464883 | 0.913043 | 0.418060 | 0.899281 | 0.984926 | -0.006398 | 0.863354 | -0.027805 |
| 11 | (Action, Drama) | (Thriller) | 0.543478 | 0.628763 | 0.418060 | 0.769231 | 1.223404 | 0.076341 | 1.608696 | 0.400000 |
| 12 | (Thriller, Comedy) | (Drama) | 0.396321 | 0.913043 | 0.359532 | 0.907173 | 0.993570 | -0.002327 | 0.936759 | -0.010606 |
| 13 | (Crime, Thriller) | (Drama) | 0.382943 | 0.913043 | 0.371237 | 0.969432 | 1.061759 | 0.021594 | 2.844720 | 0.094265 |
| 14 | (Crime, Drama) | (Thriller) | 0.453177 | 0.628763 | 0.371237 | 0.819188 | 1.302858 | 0.086297 | 2.053170 | 0.425104 |
| 15 | (Crime) | (Thriller, Drama) | 0.469900 | 0.576923 | 0.371237 | 0.790036 | 1.369395 | 0.100141 | 2.014993 | 0.508867 |

The following results were achieved by setting the minimum support to 0.35 and minimum confidence to 0.75.

We can see that these association rules are relatively similar to the association rules discovered in the first experiment. That infers that people tend to like movies with genres similar to other liked genres. On the other hand there are a few "surprises", such as **Action and Adventure -> Drama**. That example represents genres which are not really similar to each other and did not appear in the first experiment, that means that given genres should be put together, as they are usually liked by the public, even though they are different from usual genres that given public likes to watch.

# Conclusions

The association rules discovered by us provided us with additional knowledge about movie themes and their associations.

There were rules which were obvious and could be inferred without doing the experiments. These rules we have discovered in the first experiment, where similarity or affiliation between genres can be easily noticed. Not a lot of insight was also provided in the third experiment, where we tried to associate the score to a movie genre. In reality we could only see, that on average each movie theme has a rating around 3, with only two exceptions being action movies that are the only ones that can be associated with rating lower than 3.5, and drama movies that could be associated with rating better than 3, as the only one. It implies that the amount of action movies made is large, thus making it harder for the association rule of better score is difficult to achieve. Contrary to that, drama movies are rarer and appear to have a better rating, making it easier for the association rule to appear. That proves that with quantity, quality tends to drop.

Our experiments were also able to produce some surprising association rules in experiment 2 and 4, where we tested the titles‘ association to genres and the association between different genres liked by a single cinema goer. These rules, even though not apparent, had relatively high confidence and support, meaning that these rules truly have some endorsement in reality. With that knowledge, we have a different perspective on how the movie industry works, as it clearly follows association rules, even those which appeared to be surprising. With further analysis an example of **Action and Adventure -> Drama** proves to also be true and multiple movies can be found with such themes, which allows us to conclude that even some more complex associations are used in the film industry. **movie -> Comedy** or **II -> Action** were also rather unusual, but have clear reference in reality, as previously mentioned popular comedy “Scary movie” or a very popular action film “Terminator 2”.

To sum up, association rules are a great way to summarize and reflect reality in a useful and insightful way. With great parameters, interesting and useful association rules can be “mined”, which can be used in countless companies/fields/economies. They allow to develop given field in a way most appropriate to the public’s demand.