Obraz zawierający fajerwerki, ciemność, miejsce parkingowe/przestrzeń

Opis wygenerowany automatycznie

Project on Data Analysis and Mining 2023/2024

Movies

in Statistical Terms

a report in which we present selected statistical techniques  
conducted on a movies dataset

Kinga Kaczorowska  
Robert Glück

# Introduction

Movies have a strong place in our culture, we are so used to seeing various captivating stories on screens that our day-to-day world would certainly feel incomplete without them. They are not only a great source of entertainment such as action, thriller or romance films, but also information, considering for example documentary or biographical genres. We both enjoy spending time with a good movie and that is why immediately after encountering the dataset ‘Conventional and Social Media Movies’[1][[1]](#endnote-2) we were very curious to explore it further. An extra reason was that the movies collected are quite recent, dating from 2014 to 2015, so some of the titles we even saw in person, when they were displayed in cinemas.

The dataset was prepared by Ahmed Mehreev from NUST for a study trying to predict movies popularity using machine learning algorithms. More information on the topic can be found in the article [2].[[2]](#endnote-3) The initial dataset consists of 231 rows (entries) and 14 columns (features). Each movie comes with following feature names:

1. Movie: the title of the movie
2. Year: year of the premiere of the movie
3. Ratings: these values are collected from the IMDB platform. The ratings scores range from 1 to 10.
4. Genre: it is an integer categorical variable from 1 to 19 representing the genre of the movie e.g. Action, Comedy, Drama. However, the author of the dataset didn’t provide information about mapping genres and numbers so after manually checking some entries we came up with a possible classification:

* 1 = Action
* 2 = Adventure
* 3 = Drama
* 4 = Crime/Historical Fiction
* 6 = Action/ Adventure/ Comedy/Thriller/ Mystery/ Romance
* 7 = Thriller/Horror
* 8 = Comedy
* 9 = Biography
* 10 = Mystery
* 12 = Science fiction
* 15 = Horror

The rest of the numbers didn’t appear in the dataset.

1. Gross: Gross world-wide income of a paricular movie, in US dollars (collected from IMDB)
2. Budget: The total amount of money spent by creators (collected from IMDB as well)
3. Screens: Number of screens on which movie was initially launched in US
4. Sequels: an integer variable if equal to 1 it tells us the movie is either an individual or a first part of a series. Values greater than one provide information which sequel of a series a given movie is.
5. Sentiment: A Sentiment Score. 0 represents neutral sentiment; “+”sign shows the positive sentiment and the value shows its magnitude; similarly for the “–”sign and the negative sentiment. The sentiment score was calculated by Ahmed [2] through analysing the sentiments of tweets.
6. Views: the number of views represents the number of views of a trailer of movies on YouTube
7. Likes: number of likes of trailers on YouTube
8. Dislikes: number of dislikes of trailers on YouTube
9. Comments: number of comments of a trailer of movie on YouTube.
10. Aggregate Actor Followers: Number of followers of actors in one movie on twitter, only the top 3 numbers are considered.

With such data there are a lot of possibilities to train models to predict desired quantities or to explore the given information. We can try to predict Ratings and Gross, basing on the other variables, as they are the first thing that comes to mind when speaking ‘if a movie is good’ or ‘did it perform well’. Next topics to work on could be to see if the sequels do better than the original movie and if the sequel trailers behave the same, how much Likes coincide with the revenue or Actor Followers, do ratings of some Genres are usually higher than others or how much Budget and Gross are related (some financial managers would be very intrigued). Another interesting analysis would be to examine the relationship between Genre and Gross, Budget or Aggregate Followers to see which types of movies were the most trendy 10 years ago or what type of genre usually comes with a higher production costs.  
There are a lot of questions raised above, for some of them we may have an initial guess but a deepened analysis is what could give us a greater insight. In our opinion, each of the topics mentioned is a good possibility for a further study.

Throughout this report we are going to rely on not only on the well known linear regression analysis and statistical inference but also on the more advanced technique namely fuzzy c-means. A more detailed description of the methods used include: Least Squares Estimators, evaluating the normality of regression residuals, log-transformation of the data, Student’s T-test, confidence and prediction intervals, PCA and SVD decomposition and Fuzzy c-Means.

# Preprocessing the data

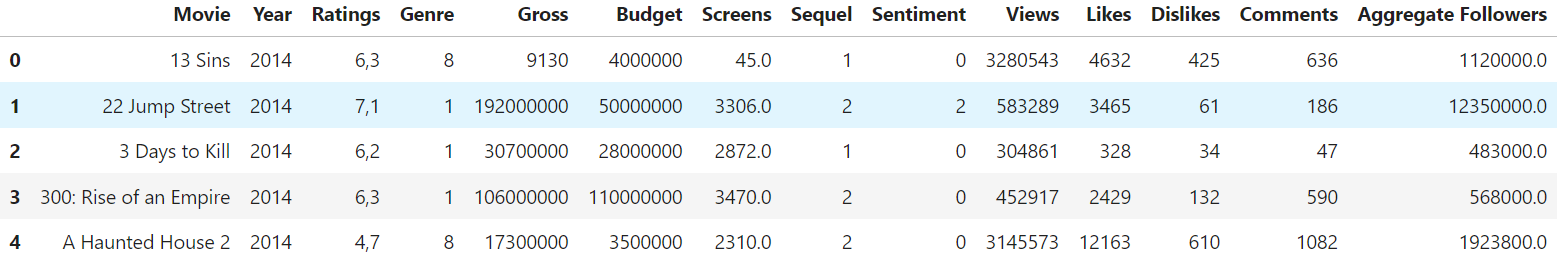
In this project we will be working only with the entries that have all of the features described above, that means we will be ignoring movies with missing values. After removal of the missing values our dataset consists of 187 rows and the number of columns hasn’t obviously changed.

Figure 1: First look at the data

It is now essential to check the data types in the dataframe to ensure that the operations we are going to perform will be properly executed.

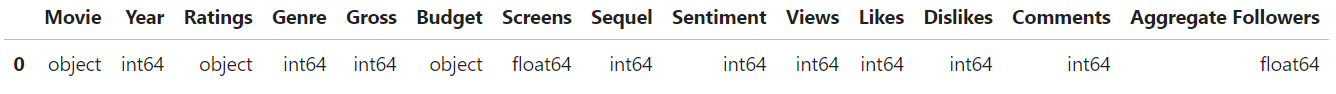


Figure 2: Features' data types

As we can see, some features are not assigned to a proper datatype. For example clearly the ‘Budget’ column should not be formatted as *object* but as the *float* number instead. Regarding all of the data types, the following adjustments are needed:

* + 'Movie', as a name of the movie, can have a type *string*
  + 'Ratings' column contains numbers from 0 to 10, usually with some decimal points, so should have type *float*
  + 'Budget' should have a *float* type as well
  + 'Screens' can be reduced to *int* since the number of movie projections is always an integer
  + 'Aggregate Followers' will be reduced to *int* likewise. There can't be 0.4 of a follower on actor’s social media

There are also categorical features such as 'Year' and 'Genre' that are currently assinged a numerical type. It is just a minor change but for consistency we are going to change their type to *categorical.*

Moreover, we noticed the decimal point in 'Ratings' column is expressed with a comma ',*00'* instead of the standard *'.00'*. For simplicity and consistency with the usual format *'.00'* we are going to ensure every *float* is expressed that way.

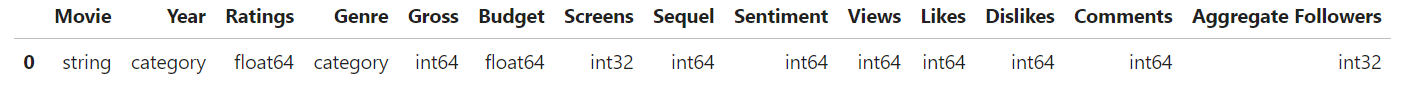


Figure 3: Data types after conversion

# Exploratory Data Analysis

With the proper data types we can begin to explore some information that could give us an insight into the data. First, let’s compute the very basic statistics such as Min, Max, Mean, Standard Deviation and Median. From that we will be able to notice each feature’s magnitue and compare a bit their distributions.

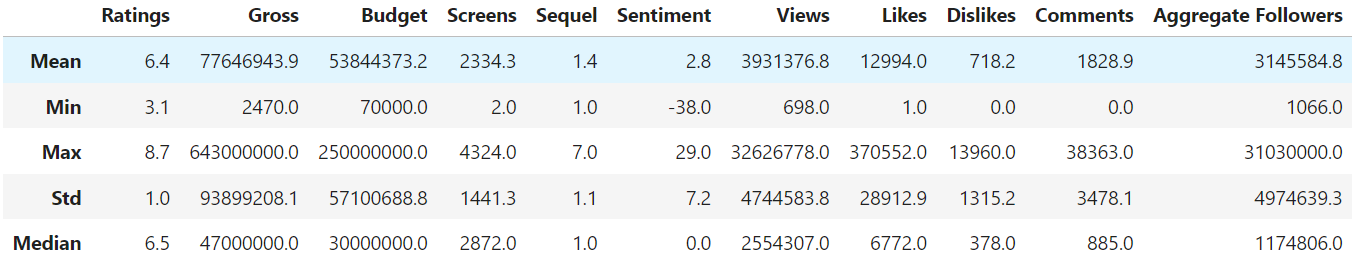


Figure 4: Basic statistics regarding each feature

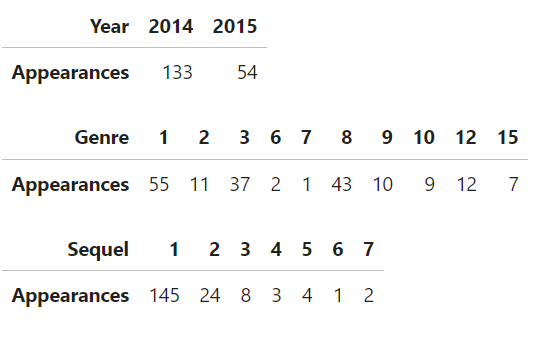
When it comes to the variable that shouldn’t have many possible values such as 'Year', 'Genre' and 'Sequel', we can get more information by looking at all the possible values and count their occurrences.

Figure 5: Year, Genre, Sequel values and appearances

The Years of the movies’ premieres aren’t equally distributed between 2014 and 2015. We are working with almost three times more films from 2014 than 2015.

The most prominent genres are 1, 3 and 8 which will probably mean Action, Drama and Comedy respectively. On the contrary from the genre 7 we only have 1 entry.

As one could have expected most of the films analysed are an individual film without any sequels and that is why the number 1 in the ‘Sequel’ column appears so often. Then the higher we go with the sequel number the less likely it is to encounter such a film.

Heading now to the most important part of exploratory data analysis, let’s look at the features’ pairplot where we can find the distributions of each values presented as histograms on the diagonal as well as scatterplots between every combination of two features.

Obraz zawierający tekst, zrzut ekranu, pismo odręczne, wzór

Opis wygenerowany automatycznie

Figure 6: Features'  
 pairplot

We are looking for a pair of variables to perform a linear regression analysis on. They should be correlated with each other enough so this analysis would make sense. What stands out when looking at the above pair plot are the combinations of ‘Likes’ and ‘Views and ‘Comments’ and ‘Views’ which correspond to the row 8 column 7 and row 10 column 7. Let’s see the correlations between these features and compare them with the rest of the dataset.

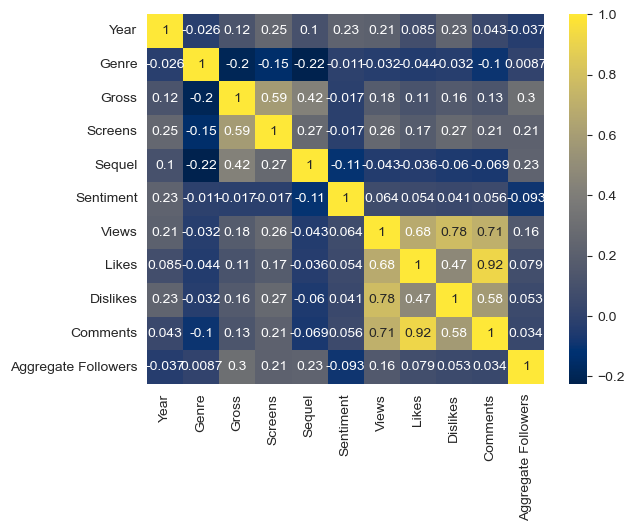


Figure 7: Correlation coefficients between features

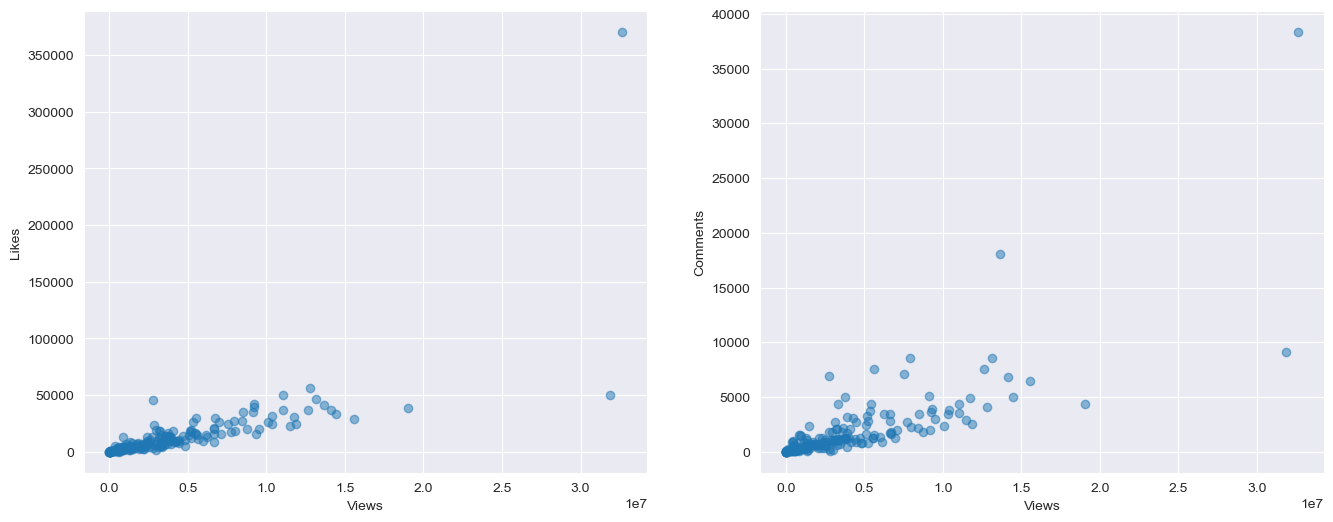
We can see the combinations mentioned have correlation coefficients equal to 0.68 and 0.71 respectively. In order to choose only two variables for linear regression, we will compare the scatter plots side-by-side.

Figure 8: Scatterplot between 'Likes' and 'Views' (left), Scatterplot between 'Comments' and 'Views' (right)

The linear relationship between the variables seems stronger in the first scatterplot. There is only one big outlier (near the value 30M Views). Possibly another point with similar number of views can be regarder as an outlier but it isn't that clear without drawing the regression line.  
We can also see that the smaller our independent variable gets, the relationship appears to be more linear in both cases.

We can try to examine this observation by plotting what percentage of the number of Views is the number of Likes for a particular movie. That should give us more insight into the area near the origin of the previous plot as well as reduce the impact of the outlier that stands out (which is a highly viewed and well recieved movie). We suspect the points on the plot of Likes/Views to Views may accumulate in an interval around a horizontal line.

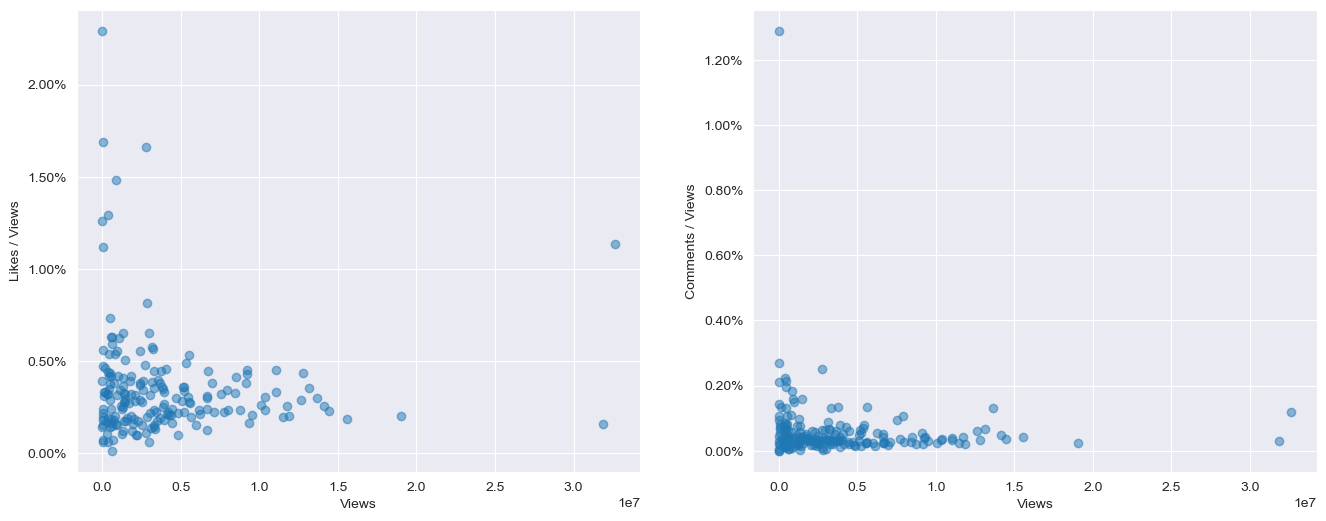


Figure 9: Percentage of Likes per View (left), Percentage of Comments per View (right)

Most of the data from the first plot follows our claim, setting the Likes interval from 0.1% to 0.5% of the Viewership. The mean of the percentage seems to be around 0.25% but we'll assess this value better during regression analysis which we are going to perform soon.

The situation on the second plot is much more complicated since the points accumulate strongly near the x=0 line so we should make this plot in a different scale, ignoring the outlier that interferes with the y-axis units. However, in this case, it would be best to estimate the slope of the regression line to get the desired value.

For the Linear Regression Analysis we decided to choose variables ‘Likes’ and ‘Views’ (scatterplots on the left side).

# Linear Regression between Likes and Views

Computing the Least Squares Estimates of the parameters of the regression line yields the following result:

Obraz zawierający zrzut ekranu, tekst, Wykres, linia

Opis wygenerowany automatycznieObraz zawierający tekst, Czcionka, biały, paragon

Opis wygenerowany automatycznie

Figure 10: Fitting regression  
line for Likes and Views

In our case the intercept has no interpretation since there isn't a movie with 0 Views (which we are going to check later). However, the slope of the regression line has its meaning; for each increase of 1000 Views, we estimate the number of Likes to go up approximately by 4.4.

The coefficient of determination is equal to 0.52 which means that 52% of the variance of ‘Likes’ is explained by the ‘Views’ variable. The value of 52% isn’t so high but we lack any comparison to a similar case.

Performing the regression analysis comes with some underlying assumptions about residuals and it’s crucial to check if they are valid in our case. We’ll do it by looking at the both normality plot of the standardized residuals and the scatterplot between standardized residuals and fitted values.

Obraz zawierający zrzut ekranu, Wykres, linia, tekst

Opis wygenerowany automatycznie

Figure 11: Checking regression assumptions

The Q-Q plot indicates that the error terms probably follow a left-skewed normal distribution but we can't see the pattern there clearly due to the scale of the y-axis affected by the outliers. Let's make the plots again, this time ignoring the big outliers on the Q-Q plot in order to get a clearer picture for further analysis.

Obraz zawierający zrzut ekranu, Wykres, linia, diagram

Opis wygenerowany automatycznie

Figure 12: Checking regression assumptions (rescaled)

The 'Q-Q plot' depicts some deviation from linearity which means that the residuals don't follow exactly a normal distribution. However, we can claim that their distribution is similar to normal but with some left-skewness and a shorter right tail.

Another important thing to notice is the accumulation of the points around the median/mean. The residuals may well accumulate more densely around the mean than a normal distribution but we can't conclude it from this Q-Q plot. Further insights can be extracted by looking at the second plot; it should make us suspect the residuals violate the following regression assumptions:

* *Zero Mean Assumption* (Because for very small x-values the residuals are positive and for the larger the x-value, the residuals tend to take more negative values. If this assumption was satisfied, the error terms should oscilate around 0, regardless of x-value)
* *Constant Variance Assumption* (Residual variance increases as the x-values increase, contradicting the assumption that the variance of the error terms is constant)
* *Independence Assumption* (It is violated specifically for the small values of x-axis; we can see there is an aggregation of points with positive value that are close to each other. Therefore, being given a residual, we can predict the vaules of the neighbouring ones. *Independence Assumption* means we shouldn't be able to do that as we assume these values to be independent from each other)

Obraz zawierający zrzut ekranu, diagram, Wykres, linia

Opis wygenerowany automatycznieThe left-skewness is also visible on the normalized histogram of standardized residuals.

Figure 13: Normalized histogram  
of standardized residuals

## Log-transformed data

The first step to apply log-transformation is to discard the 0 vales from the independent variable. We did that and it confirmed that there is no movie with 0 Views. The linear regression between the ln(Likes) and the ln(Views) fits much better than the previous one.

Obraz zawierający zrzut ekranu, linia, Wykres, tekst

Opis wygenerowany automatycznie

Figure 1414: Linear regression between ln(Likes) and ln(Views)  
We should notice the plot above doesn’t begin at 0 for the x-axis.

The coefficient of determination equals 0.89 this time which confirms what we can clearly see; that the linear relationship is stronger in this case. Let’s now evaluate if the regression assumptions.

Obraz zawierający zrzut ekranu, linia, Wykres, tekst

Opis wygenerowany automatycznie

Figure 1515: Checking assumptions of the regression between logarithms of variables

After applying the log-transformation to the data, the residuals satisfy the regression assumptions significantly better. From the Q-Q plot we conclude they follow nearly exactly a normal distribution. The only exeption are the tails; the left one is lighter than a normally distributed one and at the right tail there is a slight accumulation of points at some place.

Analysing the remaining assumptions on the second plot:

* + *Zero Mean Assumption* may well be satisfied since the residuals from the upper half seem to balance the ones at the bottom
  + *Constant Variance Assumption* can also hold but there is a concentration of points of smaller magnitude on the right side of the plot. In order to evaluate if the group of points disturbs the assumption, we should conduct a statitical test.
  + *Independence Assumtion*, similarly to the previous one, is better satisfied than in the regression analysis on non-transformed features. However, the concentration of points on the right side probably violates this assumption. With such ambiguities, it would be best to further assess our claims.

The shape of the normalized histogram looks more similar to the normal distribution as well. Therefore we can assume the normality of the residuals is at an acceptable level.

Obraz zawierający zrzut ekranu, Wykres, diagram, linia

Opis wygenerowany automatycznie

Figure 16: Normalized histogram  
of standardized residuals  
(log-transformation)

## Linear Regression Population Equation

Checking the regression assumptions was important because if they are true, we can say the observed samples are the realizations of a theoretical model of the linear relationship regarding the whole population:

* are both random variables corresponding to the dependent and independent variables respectively.
* is a population parameter of the intercept of regression line meaning that if we drew all of the points of the population on a scatterplot, the estimated regression line’s intercept would be equal to .
* is a population parameter too, but symbolysing the slope of the population regression line.
* denotes the error term between the value of and the . It is worth to mention that the error term is also a random variable which should satisfy the assumptions we were checking before i.e. it is normally distributed with 0 mean, the variance of the error terms is constant for each value of x and the values of are independent from each other for different values of x.

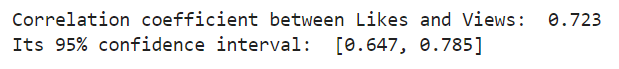
## Statistical Inference

We are going to conduct upcoming inferences on the simple linear regression between 'Likes' and 'Views', not the logarithmic one.

First we will begin by performing the T-test for the linear relationship between the two variables meaning we are going to test the hypothesis against the hypothesis (if the slope is equal to 0 there isn't any linear relationship between x and y). The confidence level of the testing is 95%.

The resulting P-value has more than eight 0’s on the first decimal places so it is much smaller than and therefore we should reject the null hypothesis which means there exists a linear relationship between features 'Likes' and 'Views'.

The next step is to construct the confidence interval for the slope of the regression line.  
  
An interpretation of the confidence interval is that if we had 100 samples and constructed a 95% confidence interval for each of them, then the true population parameter of the slope would belong to approximately 95 of these intervals. Another more loose interpretation would be that we can be 95% sure that the slope population parameter lies within the confidence interval, however, this statement is not 100% correct.

We will repeat the construction of the confidence interval for the correlation coefficient:  
  
The interpretation of the confidence interval is similar to the previous one, only substituting slope for correlation coefficient .

The confidence intervals can be also made for the y variable given x, because for a fixed value of x, y is a random variable (x is not a random variable beause we assume it is known).

First, we have fixed the x value to be the 160th in our dataset so it is equal to 3920842. We have calculated both the 95% confidence interval for the mean of random variable y for this x value and the prediction interval for the value of y given x=3920842.

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

The interpretation of these confidence intervals is basically the same as for the slope of the regression line; if we had 100 samples and given constructed a 95% confidence interval for the mean an the value of y variable for each of them, then the true population mean of the y variable given x (which is equal to ) would belong to approximately 95 of the confidence intervals and the value of would lie in more or less 95 of the calculated predictions intervals respectively.

We have also made a visualization of the 95% confidence intervals of the mean of y given x and for the value of y given x for each point in out dataset.

Obraz zawierający zrzut ekranu, tekst, linia, Wykres

Opis wygenerowany automatycznie

Figure 17: Confidence intervals for y-mean and y

# Principal Component Analysis

For the PCA we are selecting following attributes: Ratings, Views, Likes, Dislikes, Comments, and Aggregate Followers since each of them reflect how a particular movie was received by the public. They provide information on how much attention did it got and if the viewers were captured by it.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Figure 18: Selected features for the PCA

We are going to explore how the two types of normalizations perfoms; after centering the data we normalized it by ranges and by the standard deviations.

Obraz zawierający tekst, zrzut ekranu

Opis wygenerowany automatycznie

Figure 19: 2D PCA visualizations

Here we can examine the visualizations of the data projected on the first two principal components for the normalization by range as well as by standard deviation. Let's see the 3D visualizations of the projections.

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznie

Figure 20: 3D PCA visualizations

Apparently, the points take on the 3D plots take very little space, due to the perspective, which makes these plots less readable than the 2D ones. In this case, the 2D plots seem more intuitive and provide more insight into the data distribution.

## Visualization of the Selected Features using Singular Value Decomposition

Obraz zawierający tekst, zrzut ekranu, Wykres, diagram

Opis wygenerowany automatycznie

Figure 21: 2D visualizations obtained from SVD

Obraz zawierający zrzut ekranu, tekst

Opis wygenerowany automatycznie

Figure 22: 3D visualizations obtained from SVD

In every pair of plots above, for PCA as well as for SVD, we can see the normalization by range does a better job in reducing the data magnitude because the points don't overlap with each other so much as if we normalize by standard deviation. Although the absolute values of the coordinates are higher when we normalize by std, this method of scaling seems to have squished our data too much, making it impossible to see what is happening in the area where points accumulate.

The reason why the normalization by range works better could be that some features have a different magnitute than the others (for example compare Ratings with Views or Aggregate Followers) so, when projecting onto the orthogonal base, some coordinates should still remain bigger than the others. Especially because such projection corresponds to the rotation of the space. Range normalization solves this problem by rescaling each feature to the interval [0,1] so that both ends are attained by some points. Also, it's worth to mention we don't have any heavy outlier among the selected features since otherwise range normaliztion would squish the remaining points together to a very small interval (that rescaling would happen of course along the feature we observe the big differences in). Then, the normalization by std would be a better option as it's less sensitive to outliers but, having in mind our data, it isn't the case and the differeces in magnitudes play much more significant role.

The visualizations obtain by PCA and SVD look very similar to each other. Let's examine it with a side-by-side 2D plots where the x-axis and y-axis of the SVD are flipped.

Obraz zawierający zrzut ekranu, tekst

Opis wygenerowany automatycznie

Figure 23: Comparison of PCA and SVD

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Examining the differences between projections confirms our observation; for the first three principal components/right singular vectors the visualizations are essentially the same. Let's proceed further with the PCA projections

We are going to see now if there is any visible group on our visualizations regarding the genre. If so that could give us a clue that there is a relationship between selected features and a genre of the movie.

Obraz zawierający zrzut ekranu

Opis wygenerowany automatycznieObraz zawierający zrzut ekranu, tekst, diagram

Opis wygenerowany automatycznie

Figure 24: Figure 24: 2D PCA visualization with Genre labels

Figure 25: Figure 25: 3D PCA visualization with Genre labels

## Evaluating the PCA projections

We will present below a short summary of the PCA projections containing the proportion of variance captured by each principal component, its cumulative sum and the eigenvalues. Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

Figure 26: Quality of the PCA projections (by range)

Obraz zawierający tekst, diagram, zrzut ekranu, linia

Opis wygenerowany automatycznie

Figure 26: Graphical evaluation of PCA (by range)

In order to compare the normalization by range to the one by standard deviation, we will create a summary for the second option as well.

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie  
Figure 27: Quality of the PCA projections (by std)

Obraz zawierający tekst, linia, diagram, zrzut ekranu

Opis wygenerowany automatycznie

Figure 28: Quality of the PCA (by std)

In both cases the 3D visualizations consisting of the first three Principal Components should be enough to capture the essential information of the data as they retain around 90% of the total variance. When we look carefully at the cumulative sum values, we can see the 3D visualization by range is silghtly better, with a proportion of explained variance over 90%, when the std normalization attains a level close but below 90% which is ofter considered a good threshold.

When it comes to the 2D projections, their quality is basically the same for each normalization, maybe with a marginal advantage of the std. However, we have examined before that scaling by standard deviation resulted in our projections being more clustered which can make the plot less readable. That is why choosing the range normalization is a more reasonable choice in this case.

To summarize, the go-to PCA visualizations that bring the most clarity for the viewer appear to be the 2D and 3D ones obtained from the normalization by range.

# Fuzzy Clustering with Anomalous Patterns

In fuzzy clustering we are not considering columns 'Movie', 'Year', 'Genre', because they contain categorical or non-numeric data that cannot be directly used for clustering. As range of value c (number of clusters) we are testing range of values from 2 to 20. As a result of the code with fuzzy c-means (FCM) clustering for different values of c and evaluating the clustering quality using the FCM clustering criterion, we get graph of FCM clustering criterium in function of c.

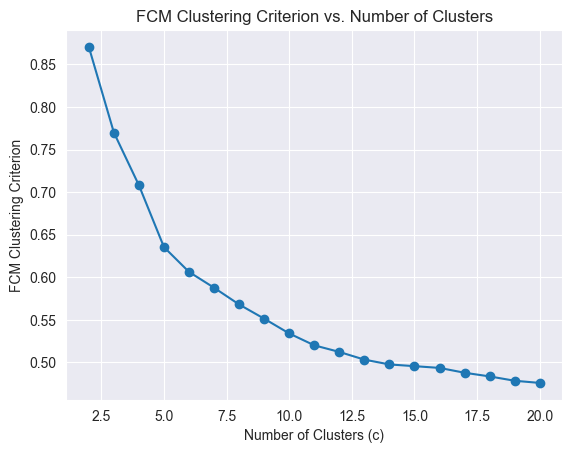


Figure 28: Clustering criterion vs Number of Clusters

It can be observed that the larger the number of clusters, the lower the value of the FCM criterion. This function decreases exponentially. This is due to the fact that as it increases the number of clusters, each data point tends to have a closer centroid, resulting in a smaller overall cost function. The values of FCM Criterion have no bigger differences after around c=10. After value c=6 the function starts to flatten out. It can be assumed the value c=5, because it stands out the most and it almost creates an angle.

## Iterative Anomalous Pattern (IAP) clustering algorithm

The Iterative Anomalous Pattern algorithm is a method used for detecting anomalous patterns within a dataset. It works by iteratively forming clusters around potential anomalies until no significant anomalies remain.

Firstly, before applying an anomalous pattern, the data were normalized the data by range, because after PC analysis in our dataset, this normalization is best. It wasn’t applied in Fuzzy algorithm, because there was no difference. Then applying the Iterative Anomalous Pattern from provided tutorial with additional changes was made. It was chosen that the threshold = 5, because then the assumed value of c fits. To visualize the data, PCA was adpated on result of applying the fuzzy-c-menas with anomalous pattern.

In the Figure 29, each navy data point is represented in a two-dimensional space defined by the first two principal components (PC1 and PC2) as it was better for our dataset, which was tested eralier in part I. The navy points represent the original dataset, while the red points represent the centroids obtained from the AP-FCM clustering. This visualization allows us to observe the distribution of data points and the clustering pattern generated by AP-FCM.

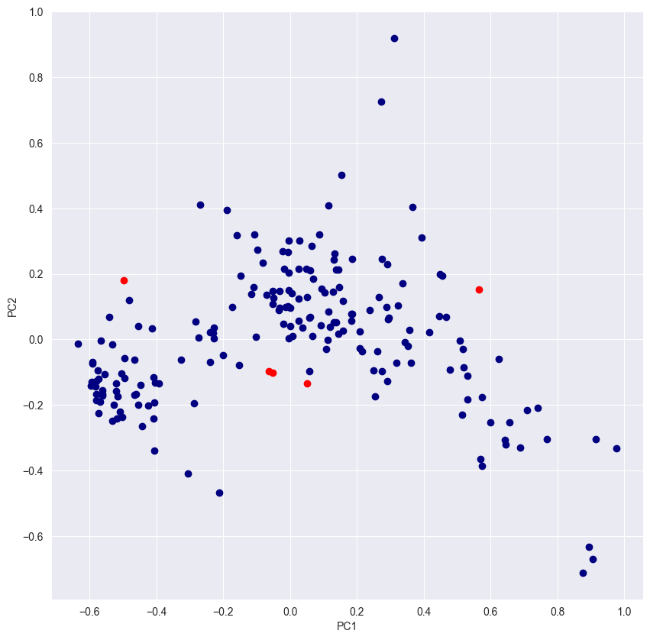


Figure 2927: Figure Plotting PCA with threshold = 5

Figure 30: Figure Plotting PCA in clour with membership to clusters

To show the membership to 5 clusters the datapoints were coloured in the plot in the Figure 30.

Summary:

The threshold parameter represents the minimum cardinality required for a cluster to be classified as anomalous. So it sets the minimum number of data points needed for a cluster to be considered significant enough to be flagged as anomalous.

With threshold 1 we have 12 clusters.

With threshold 2 we have 7 clusters.

With threshold 3 we have 6 clusters.

With threshold 4 we have 6 clusters.

With threshold 5-15 we have 5 clusters.

With threshold 16-25 we have 3 clusters.

With threshold 26-35 we have 3 clusters, but with different position.

With threshold 36-48 we have 1 cluster.

With threshold >48 the algorithm is not working.

Having 1 cluster with threshold <36-48> suggests that the algorithm has identified a major trend or anomaly in the data that requires a larger number of data points to form a cluster. With threshold > 48 the function is not working and makes an error. It shows that the stringency level has become too high, due to insufficient data points to form the required number of clusters, causing the algorithm to fail. Lower thresholds lead to finer-grained clustering, capturing smaller variations in the data and identifying more clusters, while higher thresholds result in coarser clustering, highlighting only the major trends or anomalies.

The initial prototypes in Anomalous Patterns FCM are determined by the Anomalous Pattern Algorithm. The algorithm selects the point in the dataset with the maximum distance from the centroid as the first prototype. Subsequent prototypes are chosen by iteratively splitting the dataset into clusters based on the chosen prototypes and computing new centroids until no further splitting is possible. The initial prototypes will be located at points in the dataset that are considered the furthest away from each other, as this helps in capturing the maximum variability in the data.

Applying validation indices - Partition Coefficient and Xie-Beni index

Firstly, we apllied Partition Coefficient and Xie-Beni index on Anomalous Patterns FCM.

The results are as following:

Partition Coefficient (PC): 0.003799650273083967

Xie-Beni index: 0.0008751364084378164

The Partition Coefficient measures the compactness of clusters by comparing the sum of intra-cluster distances to the sum of all distances. The Partition Coefficient is 0.0038, indicating that the clusters are relatively compact.

The Xie-Beni index measures the separation between clusters and the cohesion within clusters. A lower Xie-Beni index indicates better clustering, as it implies smaller intra-cluster distances and larger inter-cluster distances. In our data, the Xie-Beni index is 0.0009, suggesting good separation between clusters and cohesion within clusters.

Then we calculated Partition Coefficient and Xie-Beni index for fuzzy c-partitions.

The results:

Number of Clusters (c): 2, Xie-Beni Index: 0.004443030754982446, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 3, Xie-Beni Index: 0.004442972736480101, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 4, Xie-Beni Index: 0.004443241852823381, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 5, Xie-Beni Index: 0.004443048911398158, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 6, Xie-Beni Index: 0.004443179352877803, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 7, Xie-Beni Index: 0.004443009283148201, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 8, Xie-Beni Index: 0.004442860398397866, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 9, Xie-Beni Index: 0.0044431102946131265, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 10, Xie-Beni Index: 0.004442748945285978, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 11, Xie-Beni Index: 0.00444337615919075, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 12, Xie-Beni Index: 0.00444284542024087, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 13, Xie-Beni Index: 0.0044427400691436055, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 14, Xie-Beni Index: 0.004442879506903908, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 15, Xie-Beni Index: 0.00444276270469336, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 16, Xie-Beni Index: 0.004443316933488017, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 17, Xie-Beni Index: 0.004442799714773956, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 18, Xie-Beni Index: 0.004442980999907927, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 19, Xie-Beni Index: 0.0044431577399020325, Partition Coefficient: 0.002965373457810803 Number of Clusters (c): 20, Xie-Beni Index: 0.004442728084849152, Partition Coefficient: 0.002965373457810803

The Partition Coefficient obtained from the Anomalous Patterns FCM clustering (0.0038) is relatively higher than that obtained from FCM with varying numbers of clusters. The Xie-Beni Index and Partition Coefficient remain relatively stable across different numbers of clusters. This suggests that the dataset may not exhibit clear boundaries between clusters, and adding more clusters does not significantly improve the overall clustering quality.

There is also possibility that we did something wrong what may influence the results.

# Conclusions

In our dataset there are features correlated with each other on a medium level so that predicting one from the other basing on linear regression may be a loose approximation of values but definitely not a precise one. During the discussion about PCA and SVD visualization we saw that there are not appraent groups that genre would form so predicting this variable from the others may well be a hard task (but probably not impossible).

Our analysis of clustering our data suggests that the Fuzzy Clustering with Anomalous Patterns, particularly with a chosen number of clusters around 5, provides relatively stable and meaningful clustering results for our dataset. However, further validation of our methodology is recommended to ensure the accuracy and reliability of our findings.

A thing worth remembering is that our dataset is about movies, which are also a part of art, so we should interpret the predictions reasonably, trating them more like a guidelines and not rely on them too much. It is a result of much unpredictability of a movie’s success. Sometimes an actor can have a brilliant role of his life, sometimes the topic is very interesting to the audience or a movie could have a lot of budget on marketing. And the movies include a lot of information that are hard to describe and quantify such as emotions, mood, a whole picture of the camera of a character of action sequence in the plot. Maybe in the future with more advanced tools, it will be possible to capture this information within data and analyse on predict this qualities with a great accuracy.

# Bibliography

1. [1] Ahmed,Mehreen. (2017). CSM (Conventional and Social Media Movies) Dataset 2014 and 2015. UCI Machine Learning Repository.

   <https://doi.org/10.24432/C5SP5T>. [↑](#endnote-ref-2)
2. [2] Ahmed, Mahreen & Jahangir, Maham & Afzal, Hammad & Majeed, Awais & Siddiqi, Imran. (2015). Using Crowd-source based features from social media and Conventional features to predict movies popularity. 10.1109/SmartCity.2015.83.

   <https://www.researchgate.net/publication/298352830_Using_Crowd-source_based_features_from_social_media_and_Conventional_features_to_predict_movies_popularity/citation/download>

   [3] Susana Nascimento. Lecture slides of the Data Analysis and Mining course 2023/24 (2nd semester) [↑](#endnote-ref-3)