Final Essay

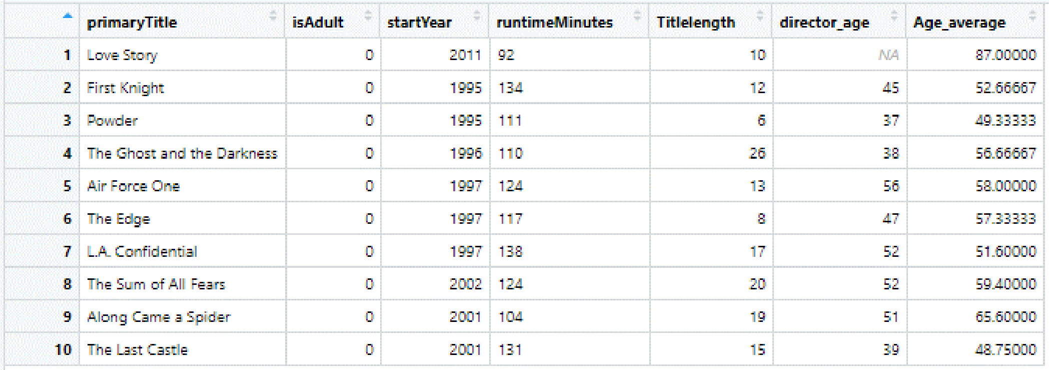
Prepared by: Nikita Gorynin
For: Construire et traiter des données avec R (niveau intermédiaire) course
By: Marine Duros

École des hautes études en sciences sociales
Paris, 2022

In the present essay, I will demonstrate the application of Principle Component Analysis (PCA) on the movie data, derived from IMDB database, using the R software with FactoMineR and factoextra packages. I will start with a description of a dataset in use, look at the correlations between the variables and finish with PCA, focusing in particular on the possible clusterization.

Data source

The data for the present essay comes from the International Movie Database (IMDB), which is a largest and most accurate online movie database, containing a wide range of information about movies, such as year of release, storyline, cast, crew, award nominations and user reviews. Originally, the data comes in a several large datasets[1], containing information separately about the movie titles, their region and language, crew and cast members and so on. I combine these separate datasets to get a table, containing a single line for each movie and some numerical variables, on which I can later exercise the PCA. These variables include the release year, marker of Adult category and run time in minutes. Few additional variables were calculated to increase the dataset, namely Titlelength (number of symbols in a title), age of the director and average age of the film crew (composer, writer, cinematographer, producer and editor), excluding the director.



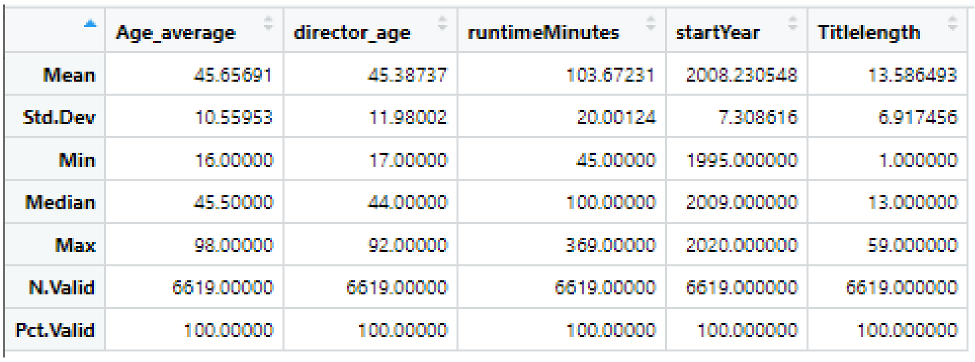
	primaryTitle	isAdult	startYear	runtimeMinutes	Titlelength	director_age	Age_average
1	Love Story	0	2011	92	10	NA	87.00000
2	First Knight	0	1995	134	12	45	52.66667
3	Powder	0	1995	111	6	37	49.33333
4	The Ghost and the Darkness	0	1996	110	26	38	56.66667
5	Air Force One	0	1997	124	13	56	58.00000
6	The Edge	0	1997	117	8	47	57.33333
7	L.A. Confidential	0	1997	138	17	52	51.60000
8	The Sum of All Fears	0	2002	124	20	52	59.40000
9	Along Came a Spider	0	2001	104	19	51	65.60000
10	The Last Castle	0	2001	131	15	39	48.75000

Figure 1. Movie data

Because the aim of PCA is to find the underlying differences, which are hard to see otherwise, I filtered the final set of movies to include only movies of one genre (Drama), from one region (US) and released within a limited scope of years (1995-2020), so that no obvious divisions can be made prior to PCA. To get rid of the obvious outliers I also excluded the movies having director or crew average ages less than zero or more than 100 years. Similarly, I excluded the movies with titles longer than 60 symbols – such titles were making a tiny minority in the dataset, but sometimes had big numbers, potentially distorting the results of the future PCA. As after all the cleansing there was only one adult movie left, I excluded this film and dropped the isAdult variable completely, to not distort the data. Finally, I excluded all the movies, for which any variables were missing to get a complete dataset with 6619 observations on 5 variables, which should be more then enough for the PCA.

[1] https://www.imdb.com/interfaces/

Major descriptive statistics for the variables in use is shown below:



	Age_average	director_age	runtimeMinutes	startYear	Titlelength
Mean	45.65691	45.38737	103.67231	2008.230548	13.586493
Std.Dev	10.55953	11.98002	20.00124	7.308616	6.917456
Min	16.00000	17.00000	45.00000	1995.000000	1.000000
Median	45.50000	44.00000	100.00000	2009.000000	13.000000
Max	98.00000	92.00000	369.00000	2020.000000	59.000000
N.Valid	6619.00000	6619.00000	6619.00000	6619.000000	6619.000000
Pct.Valid	100.00000	100.00000	100.00000	100.000000	100.000000

Figure 2. Descriptive statistics for the dataset after filtering

Directors are, on average, slightly older, then the average age within the crew, also the mean running time of the movies, although varying quite a lot, is almost equal to 100 minutes. Similarly, although the title length varies from having only 1 symbol to the maximum of 59, the mean length of the title is 14 symbols or, approximately 2-3 words.

To see the possible interrelations between the variables, I will plot a correlation matrix for them

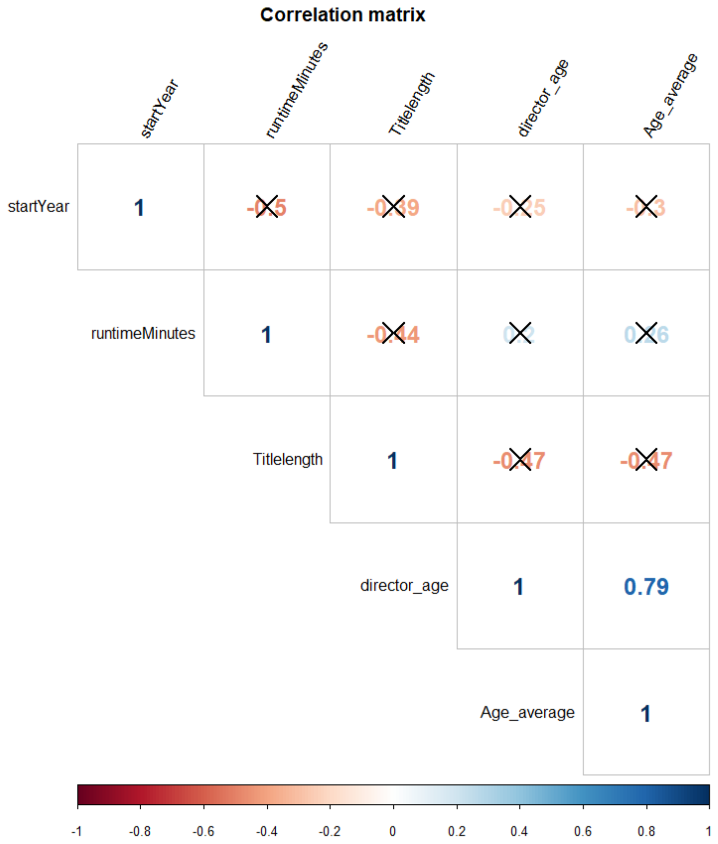


Figure 3. Correlation matrix for the variables under study

The only statistically significant correlation is found between the directors age and the crew age – this correlation is also happened to be very high (0.79), meaning that directors tend to work with people of similar age.

Principle Component Analysis

As the dataset contains 5 variables, there are 5 possible principal components to derive. To see, which of them contribute to the total variation the most, I will draw a screeplot

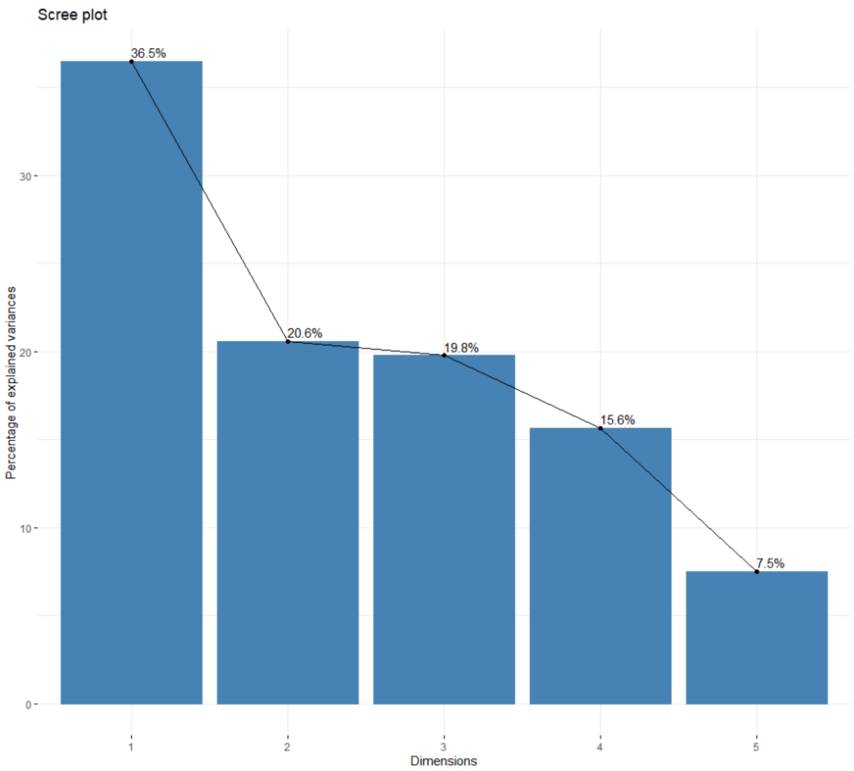


Figure 4. Percentage of explained variance by dimensions

First two dimensions capture slightly more than a half of total variation within the data: 36,5+20,6=57,1%. Thus, reducing the total dimensionality of the dataset to 2 will leave 57,1% of a total variation. Such amount is generally considered to be small, but with this essay's primarily goal being demonstrative, I will proceed with what I have.

Now, with that amount of variation left, what variables contribute to it the most?

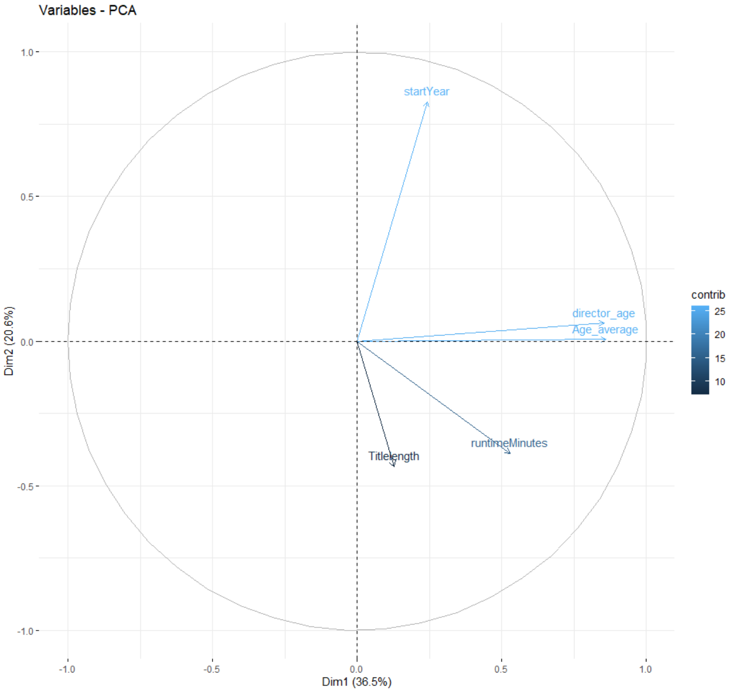


Figure 5. Contribution of variables to the dimensions 1 and 2

Clearly, the most important variables are the age of director and the crew, with both variables loading on the first dimension. The release year is also important, but mainly for the second dimension. Run time is loading on the both dimensions almost equally and title length is loading mainly on the second dimension, but weaker than the year of the release.

I will than look into individual movies on the resultant plane to check if there is any structure to be found:

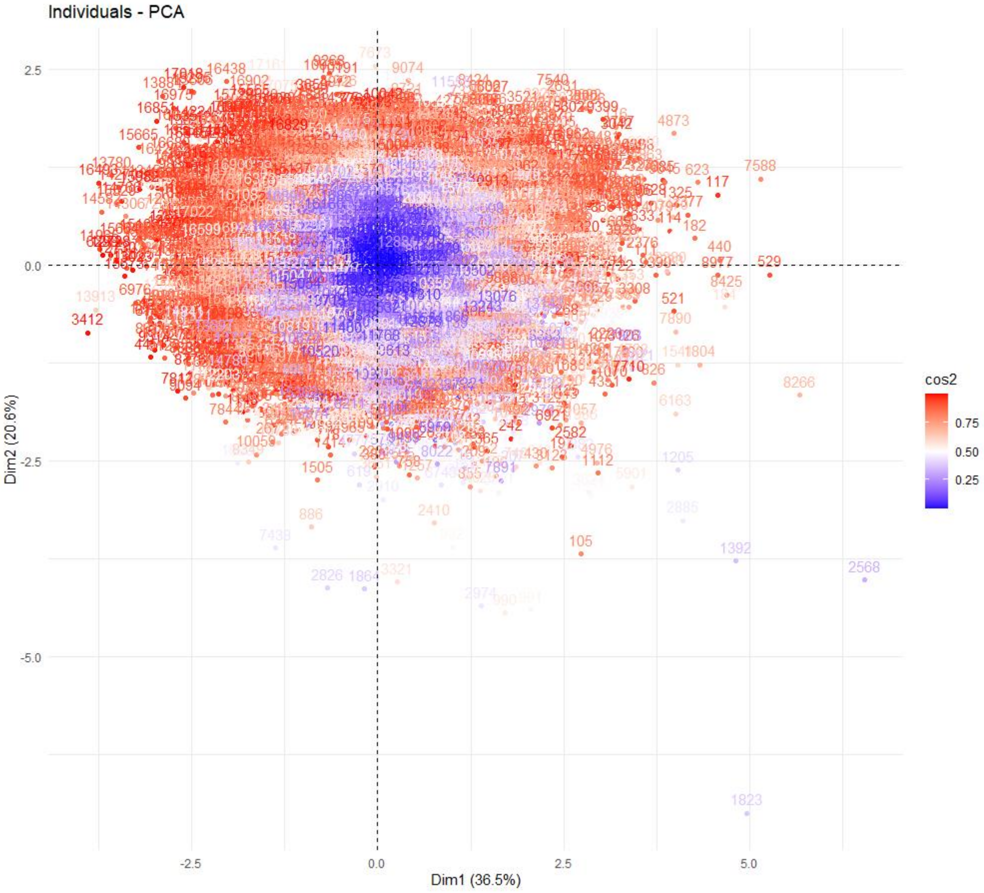


Figure 6. Contribution of the movies to the main dimensions

No visible division into groups can be found on the plot – there are few movies, which differ significantly from the majority, but overall, the structure is uniform. However, this plot is only capturing 57% of the total variance within the data, so it might be helpful to look for clusters using all the initial variables. I will do exactly this with k-means clustering and plot the result on the same principal plane as before

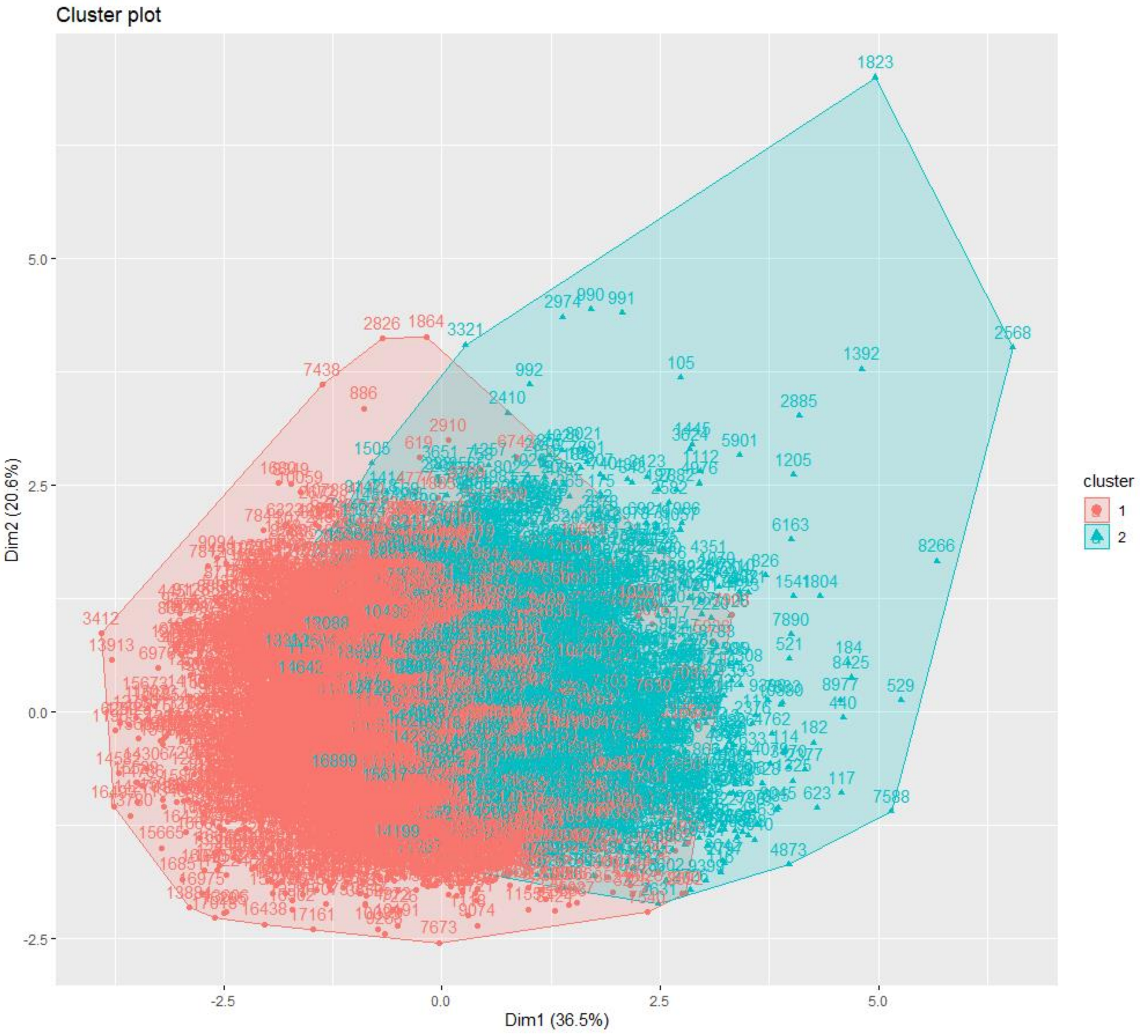


Figure 7. Structure of the movie data

Using algorithms such as k-means, it is possible to divide the data into separate clusters, however visualizing the division between them on the principal plane is not very helpful – I see that clusters protrude in one another on the principal plane, meaning that the differences are either very subtle, or they cannot be captured with only 2 principal components.

Conclusion

Movie data of 6610 movies with 5 variables, derived from IMDB database, was analyzed with Principal Component Analysis technique. It was found that age of the director and the crew are highly correlated, however no intercorrelations between them and such variables as title length, year of the release and run time were found. The variation between the data is distributed quite evenly between the variables, with first two principal components capturing only 57% of the total variance. Finally, no obvious division into groups can be made on the data, which can be displayed on the principal plane – the movies tend to have only one center on the plane.