Data understanding

1. Gathering data.

a. Outline data requirements.

First of all, we must be sure that the data we have collected for analysis is really correct, exists and is not fiction. In our case, we need to understand whether the data collected about the films in our dataset is real and true. Since they are taken from the official website of IMDB, there is no need to doubt their authenticity. All films really exist and everyone can find them on the Internet to watch, and also to find out if the information collected about them corresponds to reality. Our data set should have nominal, ordinal, discrete or continuous data, which can guarantee the possibility of further analysis and investigation.

b. Verify data availability.

The data is taken from an open source and is publicly available on the Internet, you can find fragments of a complete dataset on the IMDB website itself, but a completely and completely huge dataset for almost 7 thousand films is now available on the Kaggle platform and anyone can watch it and even experiment with it. analysis of the data in it. You can find similar data, but in a smaller volume, as already mentioned, on Kaggle or on the IMDB website itself.

c. Define selection criteria.

The main source of data is Kaggle (as was mentioned above):
https://www.kaggle.com/datasets/danielgrijalvas/movies

When we will analyze dataset, we have to understand film certificate (It will be more clear during presentation, I will give answer why it is so important to check certificate).

Motion Picture Association film rating system:
https://en.wikipedia.org/wiki/Motion_Picture_Association_film_rating_system

Central Board of Film Certification:
https://en.wikipedia.org/wiki/Central_Board_of_Film_Certification

The TV Parental Guidelines system:

https://rating-system.fandom.com/wiki/TV_Parental_Guidelines

Film certification guidance:

https://mwldan.co.uk/about-us/film-certificate-guidance

Output table from this links:

Film certificate	My certificate
UA, Passed, TV-PG, PG-13, 16, PG	middle audience
TV-14, U, G, GP, Approved	universal audience
R, Unrated, TV-MA, A, rating_NC-17, X	adult audience

2. Describing data

We have 7668 movies and different information about this movies.
This dataset consists of such features:

Name - name of the movie.

Rating - who can watch this movie.

Genre - groups of feature films, distinguished on the basis of similar features of their internal structure.

Year - released year.

Released - released date with additional information.

Score - IMDB score.

Votes - number of votes.

Director - director who was produced film.

Writer - person who wrote history.

Star - the main actor of the movie.

Budget - quantity of money spent to the film.

Gross - quantity of money movie earned.

Company - company that produced the movie.

Runtime - duration of the movie

We assume that our date should contain nominal, ordinal, discrete, continuous values. Let's check it out:

```
year               int64
score            float64
votes            float64
budget           float64
gross            float64
runtime          float64
Rate_M_A           uint8
Rate_U_A           uint8
Rate_A_A           uint8
genre_Action       uint8
genre_Adventure    uint8
genre_Animation    uint8
genre_Biography    uint8
genre_Comedy       uint8
genre_Crime        uint8
genre_Drama        uint8
genre_Family       uint8
genre_Fantasy      uint8
genre_History      uint8
genre_Horror       uint8
genre_Music        uint8
genre_Musical      uint8
genre_Mystery      uint8
genre_Romance      uint8
genre_Sci-Fi       uint8
genre_Sport        uint8
genre_Thriller     uint8
genre_Western      uint8
dtype: object
```

(We are right about that!)

3. Exploring data

In our dataset header looks like this:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000.0	46998772.0	Warner Bros.	146.0
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4500000.0	58853106.0	Columbia Pictures	104.0
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	18000000.0	538375067.0	Lucasfilm	124.0
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0	83453539.0	Paramount Pictures	88.0
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000.0	39846344.0	Orion Pictures	98.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

As we can see, our data look almost perfect! But there is some miss information in columns in the end of dataset:

...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7660	Aloha Surf Hotel	NaN	Comedy	2020	November 5, 2020 (United States)	7.1	14.0	Stefan C. Schaefer	Stefan C. Schaefer	Augie Tulba	United States	NaN	NaN	Abominable Pictures	90.0
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7000.0	NaN	NaN	90.0
7664	Dream Round	NaN	Comedy	2020	February 7, 2020 (United States)	4.7	36.0	Dusty Dukatz	Lisa Huston	Michael Saquella	United States	NaN	NaN	Cactus Blue Entertainment	90.0
7665	Saving Mbango	NaN	Drama	2020	April 27, 2020 (Cameroon)	5.7	29.0	Nkanya Nkwai	Lynno Lovert	Onyama Laura	United States	58750.0	NaN	Embi Productions	NaN
7667	Tee em el	NaN	Horror	2020	August 19, 2020 (United States)	5.7	7.0	Pereko Mosia	Pereko Mosia	Siyabonga Mabaso	South Africa	NaN	NaN	PK 65 Films	102.0

7665 rows × 15 columns

But it is normal and even such movies can be used for creating model for prediction. BUT! It should be mentioned that movies with Nan score must not be used, obviously because it is have no séance to analyze something without real score (feature that we will try to predict). It is simple to take first firsts steps and start analyze dataset, for example a lot of movies have Nan point in column “Gross”. If we analyze some of these movies, we will understand that a lot of them was published many years ago and it is not possible to find out gross, but there are some cases with modern movies and there are different reasons why it happened (for homework we think it is enough, lets save some topics for presentation 😉). And it is important to say, that we will try to use all information from dataset, but it is now clear that some features could not be used (star, director, name etc.).

4. Verifying data quality

As was mentioned above, data is perfect and there are no problems on the first look that can mess up our calculations and analysis. All data are freely available and accordingly exist.