NetFlix

Python final project

Oren Berkovich - Osnat Blau

ABOUT THE PROJECT

Netflix (NFLX) is the dominant company in the on-demand media industry, with 167 million paying subscribers around the world.

Over the years, Netflix's output has risen to a level unmatched by any other TV network or streaming service. Netflix produced a total of 371 new original shows and movies in 2019.

Features present

Show_ID - Id of the movie/TV show.

type - Movie or TV show.

title - Title of the movie/TV show.

director - Director of the movie/TV show.

Cast - actors/actress who have acted.

country - Country the movie/TV show belongs to.

date_added - Aired dated-Released date on netflix.

release year - Original movie/TV show release date.

rating - Rating of the movie/TV show.

duration - Length of the movie.

genre - Genre of the movie.

Description - Summary of the movie.

FORGING THE DATASET

In addition to the dataset we selected, we added two additional columns:

1. ranking - from random data.
2. number_of_viewers - importing data manually from online resources, if something is missing we use random.

OUR GOAL

We will compare if the rating is affected by other features and evaluate if our predictions is accurate as a real rating.

THE PROCESS

First, we will drop NaN rows, and try to minimize the amount of 'Genres' and 'Country' to get more common ones by making them more readable.

Second, we checked which data at 'Genres', 'Country', 'Rating', and 'Duration' gives more significant meaning to the rating by counting all the data values.

Third, we checked which features have more influence on the feature ‘Rating’.

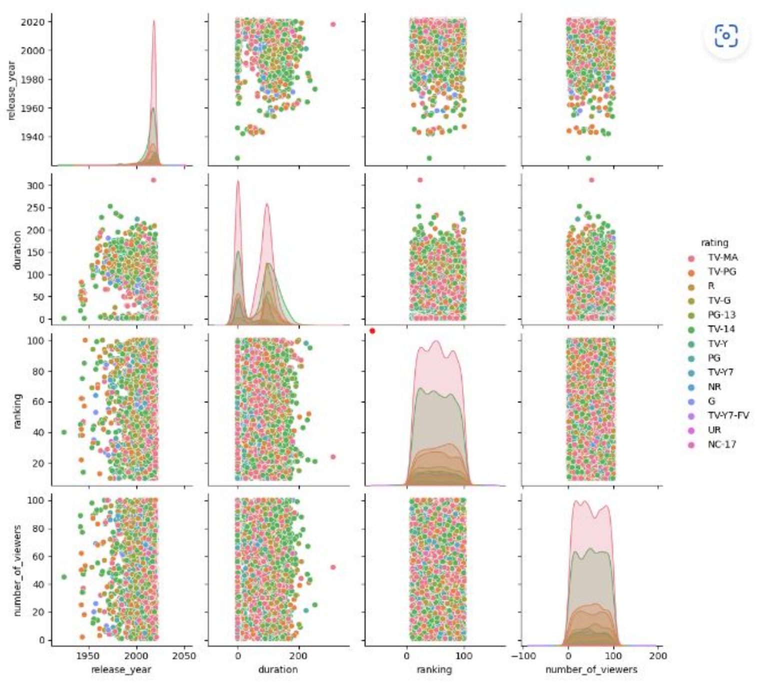
DATASET & FEATURES AFTER THE PROCESS

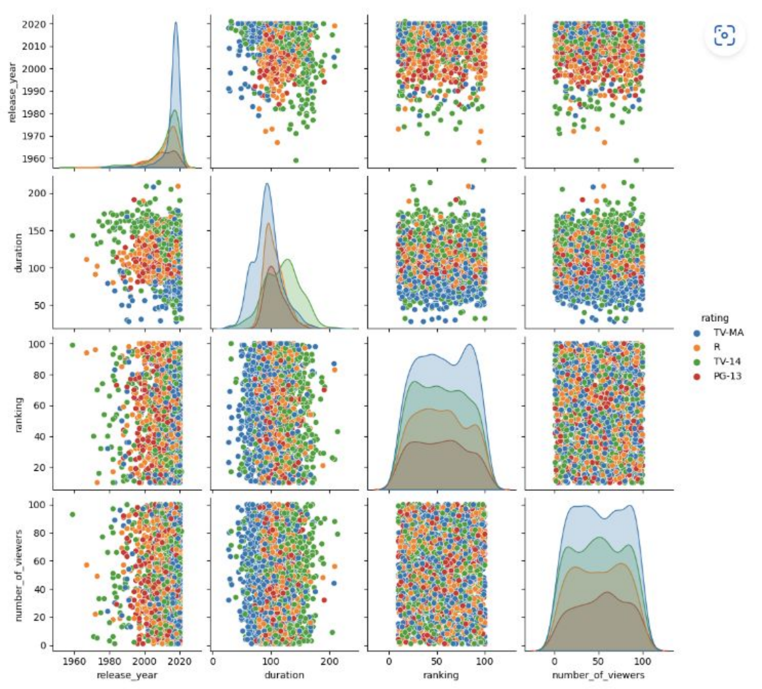
1. We removed data that could be unstructured, incomplete or inconsistent

2. We avoided a poor-data situation

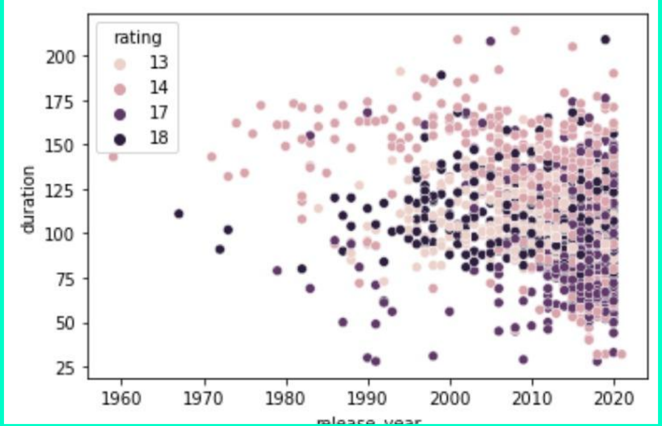
3. We reduced the size of the dataset
   from: 7787 rows × 14 columns
   to: 2378 rows × 11 columns.

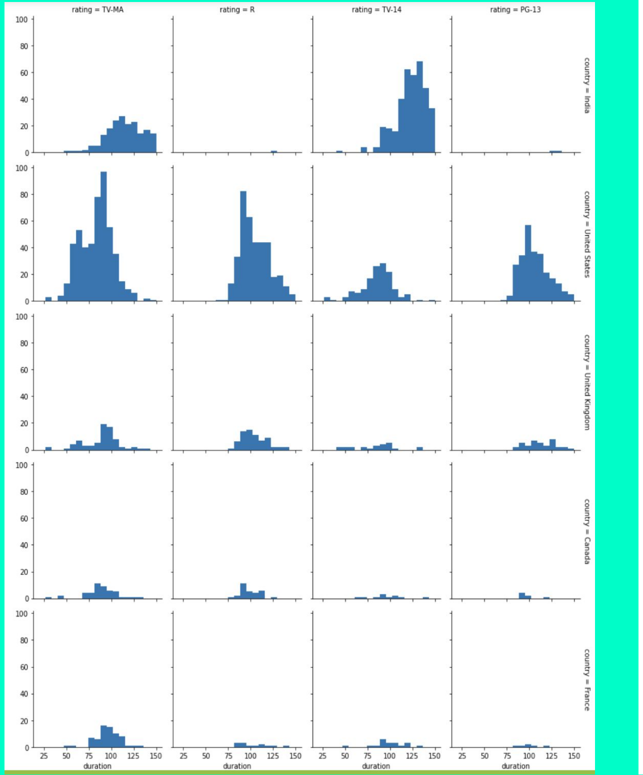
THE DISTRIBUTION OF INTERESTING FEATURES

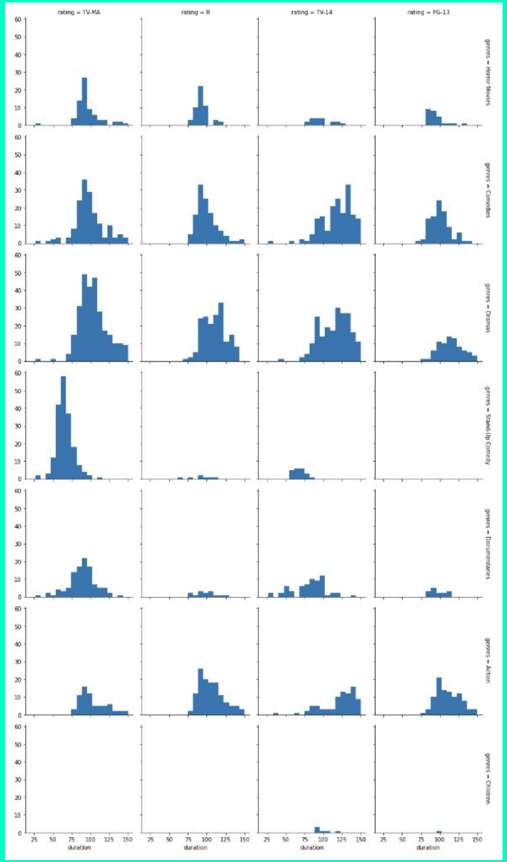




THE DISTRIBUTION OF INTERESTING FEATURES



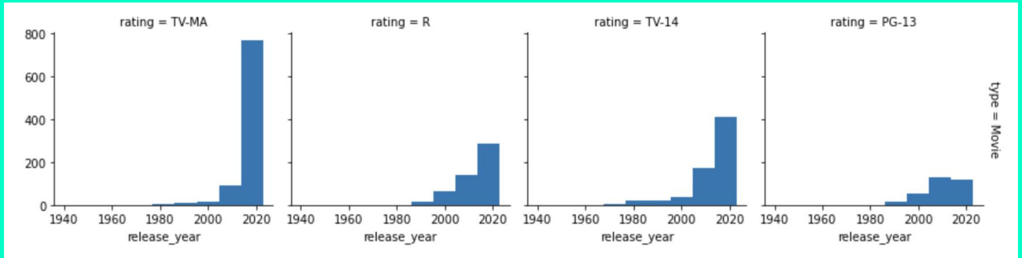




INTERESTING CORRELATIONS BETWEEN FEATURES

1. We can learn from the new dataset that Countries: USA and India are a big influence for the rating decision, and that if the movie is less than 150 minutes it has more influence.

2. We can see that Genres:
“Children” classified TV-PG, and
"Docu"/"Stand-up"/"Dramas" classified TV-MA.



THE BEST FEATURES THAT INFLUENCE THE RATING

We have come to the conclusion that the feature 'release_year' results in better dispersion and more influence on the data.

For example, between years 1980-2000 there was more PG-13 movies and in years 2000-2020 TV-14.

The process of optimizing the accuracy of the classifier model that will guess the class

We decided to use Naive Bayes classification for training and prediction as we learned in the course.

We created the model, fit him to the data, and predict the new data.

We used the bayes_plot() that visualizes a Gaussian Naive Bayes model over given data.

OUR DATA ANALYSIS PROCESS, THOUGHTS AND EXPERIENCE

- We were surprised to see the connections between the different features.
- We built a project while learning to use a dataframe in Python.
- We got more experience using Machine Learning and Classification.
- We used the different libraries that Python has to offer, such as numpy, Pandas and Matplotlib.
- We learned to work in a team and even enjoyed it.

THANK YOU!