Data Visualization

Author: Hakob Hakobyan

**Netflix** data visualization

Our [data](https://www.kaggle.com/datasets/shivamb/netflix-shows) consist of 12 columns and 8807 rows. Each row represents data of a show of Netflix including show id, type (tv-show or a movie), title, director, cast, country, date added, release year, maturity rating, duration, the genre that the show is listed in and the description.

We can do some general analysis of this data to find out the following:

* Which countries produce most of the Netflix shows,
* The longest lasting tv-shows based on how many seasons they have,
* What is the proportion of released tv-shows and movies throughout the years,
* What is the duration of a usual Netflix tv-show or a movie,
* How many shows have been released in every year, whether Netflix focuses on movies or tv-shows in the recent years.

We start from cleaning the data and for every task leaving only the necessery columns or rows. In fact we will only need show IDs, the type, the title, countries, release year and the duration columns for the given tasks.

For the first task we need to count the number of the rows for every country. Our initial data have some rows where there are several countries in a single cell, that’s why I have in those cases I have left only the first country mentioned in the cell. We also need to remove those rows that have missing values.

For the next tasks that are related to the duration of the shows, I have removed the word “min” leaving only the number of minutes for every movie. The same way I have removed the word “Seasons” so that only the number of seasons for each TV-show was left.

Besides filtering and sorting the data in many casis, in one of the tasks I have combined the duration in 10-length groups in order to construct the frequency plot.