Group: DatArtists
Hugo Casademont, Cezary Januszek, Marc Odermatt

Milestone 1

Dataset

The Movies Dataset from Kaggle
(https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset)

Problematic

Frame the general topic of your visualization and the main axis that you want to develop.

• *What am I trying to show with my visualization?*

We would like to show profiles of actors by exploring genres, rating, budget and revenue of the movies they appeared in. First, we'll create a unique colourful visualisation for each actor portraying the genre diversity of their portfolio. Secondly, we'll show the evolution of an actor's fame by plotting several metrics such as number of movies, ratings and revenue over time. Finally, we would like to show connections between actors based on the number of movies they appeared in together.

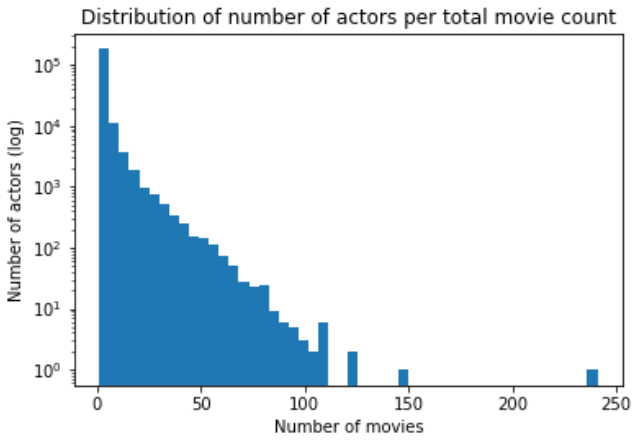
• *Think of an overview for the project, your motivation, and the target audience.*

The website should be a place for people to find out more about their favourite actors and discover new movies or related actors. The target audience is anyone interested in movies, and the interface should be very intuitive.

Exploratory Data Analysis

Pre-processing of the data set you chose
Show some basic statistics and get insights about the data

We are using the Movies Dataset which can be found on Kaggle. The dataset is composed of several csv files containing movies' metadata such as revenue and average rating, cast, crew information, keywords and individual ratings. We will mainly focus on movies' metadata and cast information.
The actors and genres were stored in a json-like structure and therefore had to be parsed. Some genres had to be eliminated as they were mistakes and were only used once.
In total, the dataset contains around 200'000 actors and about 45'000 movies for 20 different genres. We plotted the distribution of the number of actors per total movie count and found that we could reduce the number of the actors and movies by simply keeping actors who



appeared in at least 20 movies. We ended up with 3'776 actors which is more manageable while not being less relevant as we are keeping the most prominent actors. We then filtered films by keeping only those in which the selected actors appeared, resulting in 3'968 movies. Of these, we have information on 1'399 them in the dataset. In case we find older actors that are less relevant for our use case, we will remove them to further reduce the size of the dataset.
When comparing the average revenue and rating of movies before and after the pre-processing, we found that in our resulting dataset, revenue was on average thrice as much and rating slightly better which we can interpret as more viewed and appreciated films.

Related work

• *What others have already done with the data?*

Some basic data exploration projects on similar datasets have already been made. However the focus was mainly on the movies, their ratings and earnings. The cast and crew data was often simply visualised and no in-depth actor profile and connections analysis was done.

Some of examples can be found below:

1. *Exploring movie data with interactive visualisations*
   (based on the TMDB 5000 Movies dataset)
   https://towardsdatascience.com/exploring-movie-data-with-interactive-visualizations-c22e8ce5f663
2. *Data visualisation workbook on the IMDB 500 movie dataset*
   https://public.tableau.com/views/DataVisualizationonIMDB-Top5000moviedataset/DescriptiveAnalysisofIMDBMovies?%3Aembed=y&%3AshowVizHome=no&%3Ahost_url=https%3A%2F%2Fpublic.tableau.com%2F&%3Atabs=yes&%3Atoolbar=yes&%3Aanimate_transition=yes&%3Adisplay_static_image=no&%3Adisplay_spinner=no&%3Adisplay_overlay=yes&%3Adisplay_count=yes&publish=yes&%3AloadOrderID=0

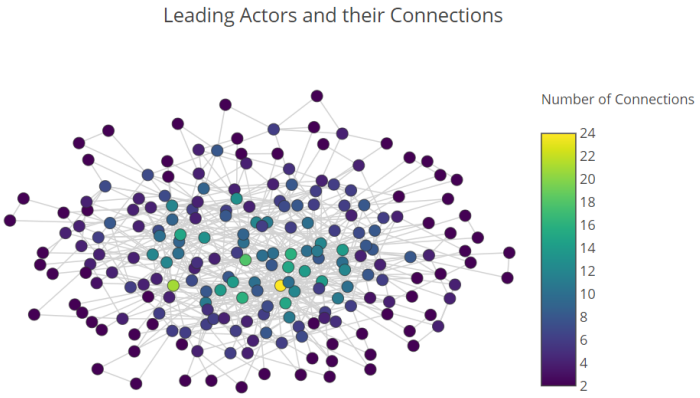
Our dataset was also often used to develop movie recommender system projects, e.g.:
Getting started with a movie recommendation system
(https://www.kaggle.com/code/ibtesama/getting-started-with-a-movie-recommendation-system)

• *Why is your approach original?*

In our project we want to focus on the actors present in our dataset. Therefore our approach is more actor-centric than movie-centric as it has been already done in other existing projects. We would like to allow people to learn more about their favourite actors and discover new movies through the actors who played in and not some other movies the user has already seen. We want to represent all the elements mentioned before in an interactive and meaningful way to allow users to discover by themselves all the information about their favourite actors they didn't know before. On top of that all these aspects will have to be visually nice and easy to understand for everyone.

• *What source of inspiration do you take? Visualisations that you found on other websites or magazines (might be unrelated to your data).*

Some other existing projects on similar datasets inspired us to choose specifically this dataset and take another approach on it. In order to create the actor connection graph we will inspire ourselves from the a simple connection graph that you can see below:



Concerning the actor's personal colour palette to represent his/her movie genre portfolio we took inspiration from an already existing website showing the user's Spotify colour palette based on the music genres he's listening to the most. Similar representations also exist for Netflix profiles.