Problem Statement - Anime Rating Case Study (Part 2)

Context

Streamist is a streaming company that streams web series and movies for a worldwide audience. Every content on their portal is rated by the viewers, and the portal also provides other information for the content like the number of people who have watched it, the number of people who want to watch it, the number of episodes, duration of an episode, etc.

They are currently focusing on the anime available in their portal, and want to identify the most important factors involved in rating an anime. You as a data scientist at Streamist are tasked with identifying the important factors and building a predictive model to predict the rating on an anime.

Objective

To preprocess the raw data, analyze it, and build a linear regression model to predict the ratings of anime.

Key Questions

1. What are the key factors influencing the rating of an anime?
2. Is there a good predictive model for the rating of an anime? What does the performance assessment look like for such a model?

Data Information

Each record in the database provides a description of an anime. A detailed data dictionary can be found below.

Data Dictionary

- title - the title of anime
- mediaType - format of publication
- eps - number of episodes (movies are considered 1 episode)
- duration - duration of an episode
- ongoing - whether it is ongoing
- startYr - year that airing started
- finishYr - year that airing finished
- sznOfRelease - the season of release (Winter, Spring, Fall)
- description - the synopsis of the plot
- studios - studios responsible for creation
- tags - tags, genres, etc.
- contentWarn - content warning
- watched - number of users that completed it
- watching - number of users that are watching it
- wantWatch - number of users that want to watch it
- dropped - number of users that dropped it before completion
- rating - average user rating
- votes - number of votes that contribute to rating

Learning Outcomes

- Data Preprocessing
- Exploratory Data Analysis
- Data Preprocessing
- Preparing the data to train a model
- Training a regression model
- Model evaluation
- Forward Feature Selection