Data understanding. C6

Data requirements

To achieve all goals stated in the "Business understanding" of the project, a dataset must contain:

1. Unique TV show ID's
2. Show names
3. Genres assigned to the show
4. First air date
5. Audience evaluation metrics: vote average, vote count

Additionally, a separate mapping between genre IDs and genre names is required to make genre-based analysis interpretable.

Verify Data Availability

The dataset is obtained from Kaggle and contains over 10,000 TV shows released from 1944 to the present.

The dataset satisfies all project requirements.

Genre identifiers were provided as numerical codes in the dataset and were resolved using The Movie Database (TMDb) API.

Initially, only the TV genre list was used, resulting in missing genre names for several shows. This issue was later resolved by retrieving the movie genre list; subsequently, all genre IDs were successfully mapped.

Define Selection Criteria

All available shows were retained for analysis, with the following filtering rules applied later in the project:

1. Shows with missing genre information were excluded from genre-based analysis.
2. Specific, very rare genres with fewer than 20 total occurrences were excluded from the "golden age" analysis to avoid unrepresentative results, for example, History, which had just five entries.

Describing Data

The dataset consists of:

1. Boolean adult flag (if show is considered Adult), not valid
2. Genre identifiers, valid
3. Country of origin, not valid
4. Original language, not valid
5. Original and translated show name, valid
6. Popularity score from TMDb, valid partially (will be explained later)
7. First air date, valid
8. Vote average and vote count, valid
9. Poster and backdrop links are not valid.

The external TMDb genre datasets (TV shows and movie lists) obtained through the API contain only two attributes: genre ID and its corresponding genre name. The final merged dataset includes a fully covered list of genres used in the primary dataset.

Exploring Data

Early exploration showed that:

1. 25% of all shows have fewer than four votes, indicating that the methodology of excluding "unreliable" ratings had to be reconsidered

2. Measuring "golden age" purely by the number of productions showed is unreliable since almost every genre burst in productions in the late 2000s

To address unreliable voting behavior, a Bayesian weighted rating was introduced to mitigate the impact of low-vote entries.

Verifying Data Quality

Overall data quality is high and sufficient for analytical purposes. The number of NaN values in columns needed for the analysis (genre, year, and weighted rating) is insignificant; for example, only 254 shows had no genre assigned.

One issue is the noticeable underrepresentation of specific genres, such as History (with only five entries) and Romance (with 14 entries).

The cause is most likely a confusing/complicated classification of several shows, where genre was pinpointed to the most precise one, missing the "secondary" genre. However, later findings confirm that a massive portion of shows have two or even more genres

The popularity rating was excluded from the golden ages since it is relevant to popularity in the present day and the recent past. According to TMDb documentation, popularity is based on the number of votes, views, and the number of users who labeled the show as a "favourite" or added it to their "watchlist" during the day. This clearly indicates that popular shows of the past century won't be comparable to even mediocre shows of today simply because they were already covered and watched many years ago, even before TMDb launched.