UNIVERSITÀ DEGLI STUDI DI TORINO

Human-Computer Interaction

Stefano Miceli, Samuele Tommasi

2024/2025

Contents

1 Introduction

This report documents the design, implementation, and results of the data analysis project focusing on movies related datasets. The project aims to extract insights from datasets covering various aspects of cinema, such as individual movie details, Oscar awards, and reviews. The analysis process involves two main phases: **Data Cleaning** and **Data Visualization**, implemented with Jupyter Notebooks.

The **Data Cleaning** phase ensures the integrity and usability of raw datasets through operations such as handling missing values, validating keys, and reformatting data. The cleaned datasets serve as a reliable foundation for the subsequent visualization phase. In the **Data Visualization** phase, exploratory and descriptive techniques are employed to uncover trends, correlations, and patterns across the datasets, despite the absence of a unique identifier linking them.

2 Data Cleaning

2.1 Solution

To ensure consistency and reliability, we adopted a standardized methodology for cleaning each dataset. This methodology defined steps for data understanding, validation, and cleaning, providing a structured and repeatable approach. By standardizing this process, we minimized oversight and enhanced the clarity of our workflow.

Specific considerations were made for each dataset, for unique characteristics and potential issues. One strategy was handling null values: invalid or unknown fields were explicitly set to `null`. This differentiation helped ensure that filled fields contained valid data and void fields were clearly marked as incomplete, to simplify visualization and database storage, and improve clarity and data integrity.

2.2 Issues

The primary challenge was identifying potential problems within the datasets. We needed to decide whether duplicate entries were permissible in this context, how null values should be handled to balance data completeness and

validity, and what dataset-specific cleaning steps were necessary and how critical they were.

For example, the runtime field in the movie dataset posed a significant challenge. Its values ranged from 1 to 70,000 minutes, with the upper extreme revealing cases of TV series where episode runtimes were stacked. This made us question whether such extreme runtimes could reasonably represent movies. We considered setting a threshold for maximum and minimum runtime values but ultimately decided against it, as movies are inherently creative works, and no definitive rule could ensure the exclusion of all non-movie entries. Instead, we opted to handle runtime inconsistencies case-by-case during database creation and data visualization.

2.3 Requirements

Our solution strictly followed the project's requirements, producing high quality cleaned data.

2.4 Limitations

The most significant limitation was our limited prior experience with Python and data manipulation techniques. As we developed the cleaning methodology, we concurrently learned the tools and best practices, which posed a steep learning curve. Despite this, the resulting solution was extensible and maintainable, enabling straightforward adaptation to new requirements or bug fixes.

3 Data Visualization

3.1 Solution

Inspired by the lecturer's advice to storytell our analysis, we approached data visualization as an opportunity to narrate meaningful stories about the data. Each visualization was designed with a clear narrative structure, starting with an introduction to the topic and context, followed by a prediction or hypothesis, the presentation of findings through visualizations, and finally a conclusion comparing results to initial predictions.

To make the analysis engaging, we differentiated both the topics explored and the types of plots used.

3.2 Issues

The most significant challenges in this phase were the need to balance creativity and technical execution. Developing unique and compelling narratives for each visualization required us to think deeply about the data and find new ways to tell impactful stories. Some visualizations, like the poster color analysis, were particularly demanding as they involved learning image processing techniques and technologies that were unfamiliar to us. This required significant research and experimentation, but ultimately enriched our analysis by adding a creative and technical dimension.

Another major limitation was the difficulty of designing clear and engaging visualizations while ensuring they remained accessible and easy to interpret. This often required multiple iterations to refine the presentation and maximize the clarity of insights.

3.3 Requirements

Our work followed strictly the project requirements by producing visualizations that were meaningful and effective. Each plot was designed to extract and communicate insights effectively and our approach to storytelling ensured the visualizations were engaging and aligned with the project's objectives.

3.4 Limitations

The lack of a shared key to link the movie, Oscar, and review datasets posed a significant limitation. Attempts to use fuzzy matching on movie titles proved unreliable due to identical titles appearing across unrelated products. This issue prevented us from exploring certain relationships, such as correlations between Oscar success, review scores, and box office performance, reducing the scope of our analysis in this regard.

4 Conclusions

The data analysis phase of the project was a valuable experience that reinforced our knowledge of data cleaning and visualization. By standardizing the cleaning process across datasets and applying specific operations for each challenge, we ensured high data quality and reliability. This enabled us to create meaningful and diverse visualizations that uncovered insights about movies, Oscars, and reviews, even within the constraints of the dataset. While the lack of a shared identifier prevented certain analyses, it also encouraged us to explore alternative methodologies and focus on the individual strengths of each dataset.

5 Division of work

In the **Data Cleaning** section, we worked together to preprocess the dataset and prepare it for analysis. For the **Data Visualization** section, we divided the visualizations equally, with each member responsible for specific charts and graphs, ensuring balanced participation and comprehensive coverage of the data. At the end of the project, all team members worked on every aspect to refine the final result across all sections.

6 Extra Information

Full instructions for running the webservice are provided in the ReadME file of the GitHub repository.