# A comedy handbook

GitHub repository

## Abstract

Laughter. This great human sentiment that surrounds us from a very early age is a crucial aspect in cinematography and even has its own genre associated – comedies. Do you remember chuckling at the latest Hangover? Do you also remember being bored when watching The Emoji Movie ? Both movies are comedies, yet one was much “better” than the other. This project aims to determine what factors make for a great comedy. We will explore four main aspects, namely what makes a movie popular, what makes it qualitative, whether there are cultural aspects influencing this, and to what extent are each of these factors unique to our genre of interest.

## Proposed plan

Hereunder, you will find the plan of the project. Each subsection will briefly outline its purpose, and how it helps in answering our leading question: how to make a good comedy. The next subsection (“Research questions and methods”) will describe, for each point of this plan, concrete questions that we will aim to answer, as well as how we plan on finding said answer.

### I. Popularity analysis

This section analyzes the factors driving comedies' popularity. We begin by defining a "popularity score," based on user reviews and gross sales. Then, we will try isolating which factors (e.g., title, plot, cast, director) mainly determine a comedy’s popularity, and to which extent they do so.

### II. Quality analysis

Similar to part I.: we will determine a comedy’s “quality score” through experts’ ratings, movie distinctions and awards, and we will then try isolating which factors determine it.

### III. Genre differences

While we do all this work for comedies, we can look at the impacts of interactions with other genres (e.g.: romantic comedies, action comedies, …) and highlight which particular factors stand out for comedies.

### IV. Cultural impact

Finally, we will attempt to contextualize all the above information by looking at whether it is bound to time (i.e.: particular historical events/social trends in the past century) and/or space (i.e.: countries, linguistic areas).

# Research questions and methods

This section outlines the specific questions our project aims to address. While some questions could apply to genres beyond comedies, we focus on comedies to extract more targeted and meaningful insights. Additionally, **Section III** will explore key differences between comedies and other genres where relevant and interesting.

## I. Popularity analysis

Other used datasets:

- Merge Rotten Tomatoes datasets (this, this, this, this)
- Missing values will be scraped off RT directly using a web scraper, similarly to this.
- Perform subset analyses on movies with complete data (drop NaNs), with a focus detailed in the final report.
- Writer information from title.crew.tsv.gz

| Question(s) and goals | Method |
|---|---|
| • Evaluate how gross sales and user reviews reflect movie popularity.<br><br>→ This will establish and justify our baseline "popularity" metric used for the subsequent questions | • Sentiment analysis on non-expert user *textual reviews* to get a user "likeability-score"[1]<br>• Correlation analysis (statistical tests) between gross sales, likeability score, and user ratings |
| • Does the title influence the popularity of a movie (e.g: title length, title composition…)? | • Correlation analysis between each title characteristic and popularity |
| • Do partnerships impact movie popularity?<br>• Do repeated collaborations improve the quality of comedy movies over time? | • To detect partnerships, identify top-K most popular N-grams of:<br>  ○ people in general<br>  ○ actors (cast) only<br>  ○ director, and actors<br>  ○ directors and writers<br>• To quantify their impact: statistical tests with the movie's popularity |
| • Do sequels have an impact on popularity? | • Detect sequels (e.g.: "Shrek 1", "Shrek 2".. can be extended using "Longest Common Substring" on movie titles)<br>• Analyze trends in popularity changes across sequels. |

[1] Sentiment analysis for reviews in datasets which don't already have it can be done using small language models, such as BERT or SLIM. If we limit the # samples, this should not take that much time on CPU as they are SLMs

## II. Quality Analysis

Other datasets:

- Movie awards using merge of this and this

| Question(s) and goals | Method |
|---|---|
| • Assess how well awards and expert reviews indicate movie quality.<br><br>→ Use this analysis to define and justify the baseline "quality" metric for subsequent questions. | • Sentiment analysis on expert *textual reviews* to get an "expert likeability-score"<br>• Correlation analysis (statistical tests) between gross sales, likeability score, and expert ratings |
| • Do specific plotlines affect a movie's perceived quality?<br>• For example, are certain themes, like war for awards, also significant for comedies? | • Clustering of plot keywords amongst plots (pre-filtering using a standards tokenization pipeline (normalization, postword removal, stemming, and, in our case, removing verbs)) |
| • According to experts, what are the key points of a good comedy? | • Clustering of review keywords for different expert *textual reviews*/consensus on "qualitative" movies. |

After Sections I and II, we will address the side question: "Who provides more valuable insights—experts or users?"

## III. Genre differences

| Question(s) and goals | Method |
|---|---|
| • How do comedies perform with other genres?<br>• Which combinations make them stand out or fail? | • Identify comedies that overlap with other genres (e.g., action-comedy, adventure-comedy).<br>• Perform statistical tests to evaluate how each genre combination performs based on popularity and quality scores. |
| • Are there comedy-specific directors ?<br>• Does the genre define the director's style? | 1. Identify each director's primary genre.<br>2. Analyze their style through common words in {expert reviews, user reviews, movie plots}.<br>3. Examine and compare style patterns. |

## IV. Cultural Impact

Other datasets :

- If we don't find a per-country per-movie, sales dataset, we can do the analysis for a few only using https://www.boxofficemojo.com/year/

| Question(s) and goals | Method |
|---|---|
| • What is the popularity of the comedy genre across various countries? | • Analyze the proportion of comedy movies produced by each country.<br>• Create a world heatmap to visualize comedy popularity globally. |
| • Are certain types of themes more successful in specific<br>  ○ Cultural contexts?<br>  ○ Countries?<br>  ○ Political regimes?<br>  ○ Languages? (dubbed, or original) | • Extract the themes in the movie description (keyword extraction+clustering on plots)<br>• Perform statistical analysis between themes and countries, dates in time, languages |
| • Are there some particular events that influenced the popularity of the comedy genre ? | • Analyze comedy popularity through time in the world and correlate with historical events<br>• Inversely, pick specific events (war, political regime change, …) and perform statistical analysis on popularity |

# Timeline

Below is the timeline for our project. Progress has already been made on some tasks, with the work documented in the group's Git repository.