The Movie Database
Or: How I Learnt to Visualize Data

Nabin Kumar Sahoo

Chennai Mathematical Institute

Abstract

This is the final report of the semester project of the Visualisation course taken during August-November, 2022 at Chennai Mathematical Institute.

The purpose of the project is to study and analyse the quantitative aspects of movies, like average rating, budget, runtime, among other variables. We also aim to study relations among different genres.

The visualisation is done in `R` using the `ggplot2` package.

1 Introduction

Cinema is, and has been since early 20th century, one of the most prevelant forms of art. The visual basis of the medium gives it a universal power of communication, and its reach is easily extended (by breaking language barriers) through the use of subtitles or dubbing.

Being a huge – pardon my presumption – movie aficionado, I can say that movies have been an essential part of my life.

1.1 Project Description

Through this project I will explore the quantitative aspects of movies – average rating (out of 10), budget, vote counts, runtime – and the relationships among them. Specifically, I will study the following relations:

- *Avg. rating vs Budget*,
- *Avg. rating vs Genre*,
- *Avg. rating vs Runtime*, and
- *Budget vs Genre.*

I will also explore the relationship among different genres. Often, as is understandable, a particular movie doesn't have only one genre attributed to it, instead there are usually two or three (or even more) genres. For example, *Quantum of Solace (2003)* has four genres, namely *Adventure*, *Action*, *Thriller*, and *Crime*. So, I will look at what (co-)genres are most common with what genres.

2 Dataset

The dataset used is the **TMDb (The Movie Database) 5000**[1] dataset, generated from TMDb API. It has data on more than 4800 movies.

There are two csv files, `tmdb_5000_credits.csv` and `tmdb_5000_movies.csv`. For now, we will solely work with the latter.

2.1 Content

The CSV file `tmdb_5000_movies.csv` has the following variables that are of interest to us:

- `id`,
- `genres`,
- `runtime`,
- `vote_average`,
- `vote_count`,
- `budget`

Other variables are of no interest to us, and we will remove them in the data cleaning process. For more detailed descriptions of the selected variables, refer to the table below.

There are other variables that could have been useful, like `original_language`, but for most (more than 90%) of the movies, the language is `en` (English).

2.2 Data Loading and Cleaning

We first remove rows with missing values. In particular, there are some entries with `runtime` set to 0, or `genres` set to `'[]'`; we remove them. We also consider only those rows that have `vote_count` $\geq$ 10.

2.2.1 Genres

The column `genres` has entries in JSON format, we need to convert it into a useable format for `R`.

Before we go into details of that, there are 20 unique genres in the dataset. These are, in alphabetical order, 'Action', 'Adventure', 'Animation', 'Comedy', 'Crime', 'Documentary', 'Drama', 'Family', 'Fantasy', 'Foreign', 'History', 'Horror', 'Music', 'Mystery', 'Romance', 'Science Fiction', 'TV Movie', 'Thriller', 'War', 'Western'.

The two least common genres (in the cleaned/filtered dataset) are `Foreign` and `TV Movie`.

There are only 4 entries with genre `TV Movie` and I wouldn't consider it a genre[2] So we remove it from the dataset.

[1] `https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata`

[2] TV is a medium, not a genre.

Variable Name	Type of Variable	Description	Example entry
`id`	Nominal categorical	Unique number to identify movies	`10764`
`genres`	Categorical	Genres of the movie (in JSON format)	`[{"id": 12, "name": "Adventure"}, {"id": 28, "name": "Action"}, {"id": 53, "name": "Thriller"}, {"id": 80, "name": "Crime"}]`
`vote_average`	Continuous numerical	Average rating of the movie across all votes	`7.2`
`vote_count`	Count	Number of votes of the movie	`19`
`runtime`	Continuous numerical	Runtime of the movie	`169`
`budget`	Continuous numerical	Budget of the movie (in USD)	`1100000`

Table 1: Variables from `tmdb_5000_movies.csv` that we will work with.

Similary, there are only 13 movies with the genre `Foreign` in the dataset, and there are way more foreign movies that are not given the genre `Foreign`, and again I wouldn't consider "Foreign" a genre[3]. So we remove that as well.

Coming back to converting `genres` from JSON to a useable formate, we do this in two ways:

1. **One-hot encoding of genres.** We create 20 new[4] columns, each corresponding to a different genre. The entry for a particular row would be `1` if the corresponding genre is in the list of genres for that movie (row), otherwise `0`.

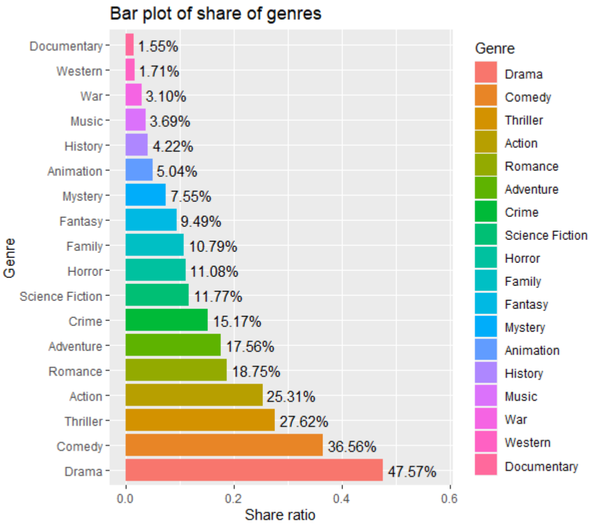
2. **Creating multiple copy of rows for different genres**. We create a new column, `genre`, and if a movie has n genres, we create n copies of that row that differ only in the column `genre` – where each genre of that movie appears once.

3 Visualization and Analysis

3.1 Genre

We first visualize the share of movies per genre. We do this using a bar graph.

For this section we will use the one hot encoding of genres.



Note that genres aren't mutually exclusive, that is, two different genres can correspond to a single movie. So the sum of percentages in the above graph will exceed 100.

From the bar graph, we can make the following conclusions:

- 'Drama' is the most common genre. Followed by 'Comedy' and 'Thriller'. This is not surprising, as these movies are all over.

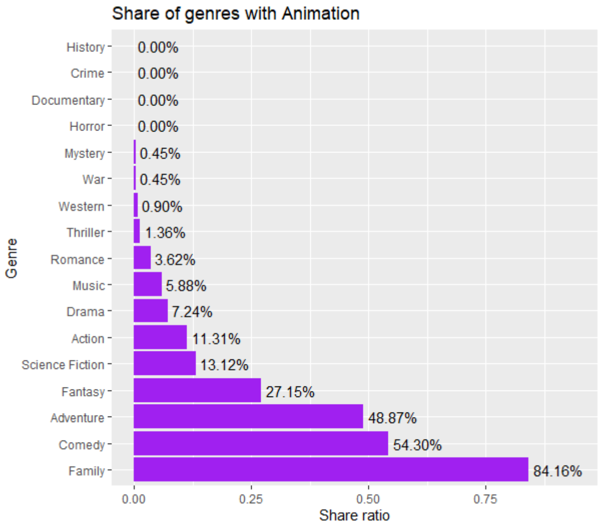
- The least common genres are 'Documentary', 'Western', 'War', 'Music', and 'History'; each have a share of less than 5%.

- A (relatively new) genre that is becoming popular nowadays is missing from the dataset, 'Slice of Life'. This is because few websites categorize movies under this genre and TMDb apparently doesn't.

Let us now look at the relationship between different genres. First, the share of movies having 'Animation' as a genre.

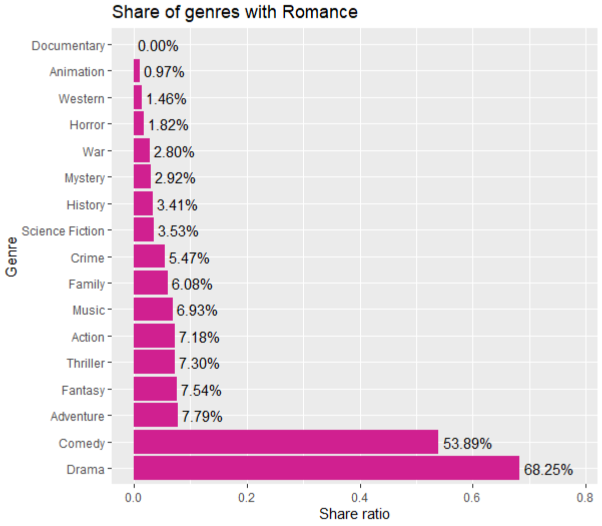
[3] The term "foreign" is subjective

[4] As there are 20 unique genres in the data.



- From this, it is clear that most of the movies under *Animation* also fall under *Family*. This makes sense, as (sadly) most of the animated movies are aimed towards kids.
- Also, note that the sum of shares of 'Family' and 'Comedy' is greater than 1. So we can conclude, that there are animated movies that fall under both 'Family' *and* 'Comedy'.
- We also see that there are no 'Animated' movies that are also 'Horror' or 'Crime'.

For one more example, we look at the bar graph for 'Romance'.



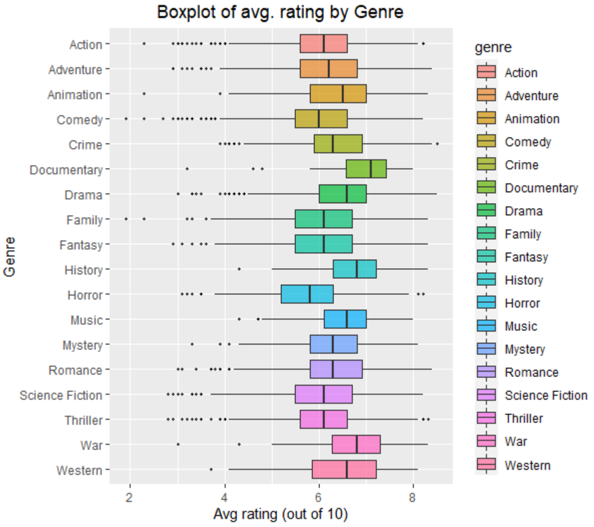
- The relationship of 'Romance' with 'Drama' and 'Comedy' is very prominent. This may be attributed to the immence popularity of *Rom-Coms*.
- There are no documentaries with genre 'Romance' – makes sense.
- Also, see that 'Animation' is more common among various genres than 'Romance' – which is only common with 'Comedy' and 'Drama', and not much with other genres. This makes sense, as unlike 'Romance', 'Animation' is not a hardcoded genre and some (including yours truly) would argue that animation is more of a medium than a genre.

Similarly, we can plot bar graphs for any genre to see its relationship with other genres.

3.2 Genre vs Rating

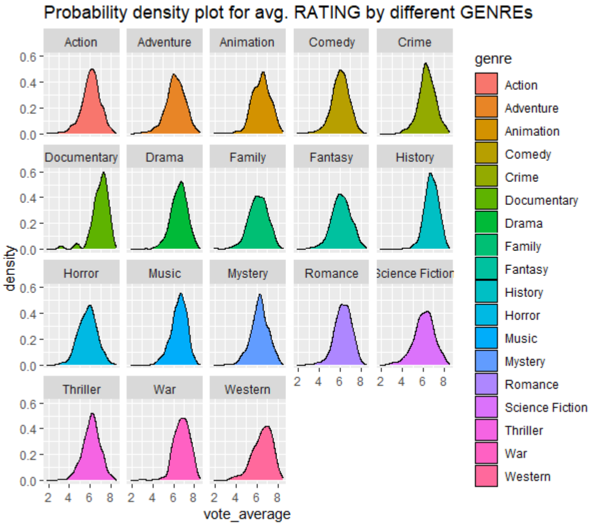
We first plot the boxplot of average rating by different genres.

Before we proceed, note that for this section we will not use one hot encoding of genres, instead, we create multiple copy of rows for different genres, as this would enable us to easily plot boxplot and density plot.



- Horror movies are generally rated lower, as compared to other genres.
- In contrast, documentaries are generally rated higher.
- 'Drama' movies are generally rated higher than 'Action' ones.

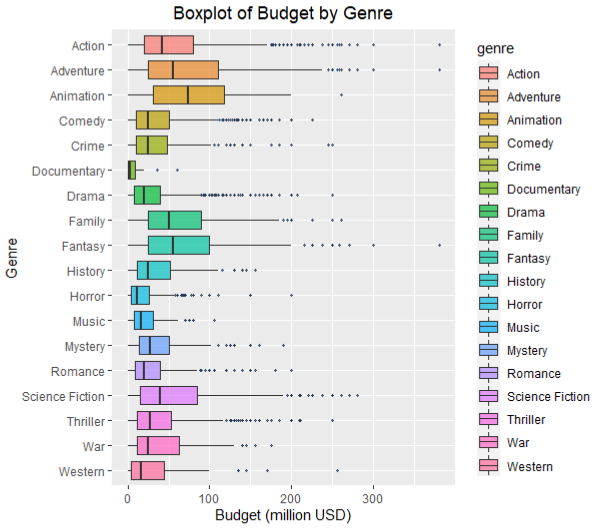
Let us now look at the probability density plot for each genre separately.



We are seeing Central Limit theorem in action.

3.3 Genre vs Budget

There are a lot of movies with `budget` entry 0, or a small positive integer. We will remove them. In fact, for this, we only consider movies with a budget more than 100,000 USD. There are still more than 3600 unique movies in the filtered dataset; this is enough for us.

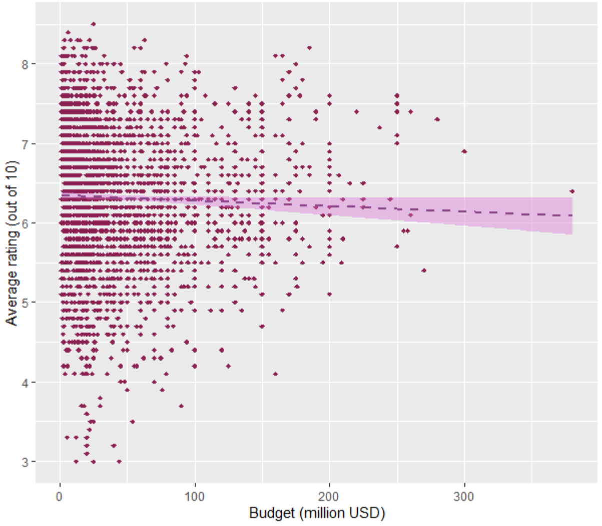


- Documentaries and horror movies tend to be cheaper than other genres.
- Drama movies are generally cheaper than Action movies but there are some high-budget drama movies as well.
- Most of the mega-budget movies fall under Action or Adventure. (Notice the outliers.)

3.4 Rating vs Budget

As in the previous section, we only consider movies with budget $\geq$ 100,000 USD.

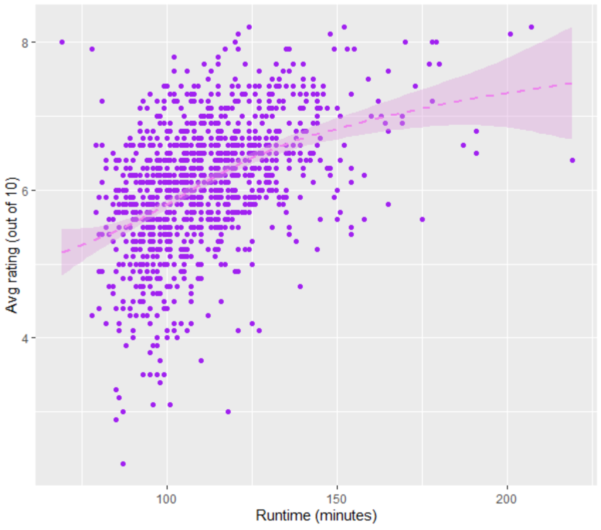
We plot the graph of rating vs budget.



- Most of the movies have a budget of less than \$100 million.
- There seems to be no apparent relationship between budget and rating.
- All the movies with a rating of less than 4 don't have budget greater than \$100 million. A very high budget somewhat ensures that the quality is not garbage.

3.5 Rating vs Runtime

We plot the graph of rating vs runtime.



- Most of the movies have a runtime $<$ 150 mins (2h30m).
- Longer movies tend to be rated higher.

4 Summary

From the entire visualisation process, the most important findings are noted below.

- It can be seen clearly that some of the genres are heavily interdependent.
- For all the genres, ratings of movies tend to follow normal distribution.
- Documentaries and horror movies are generally cheaper than other genres.
- There doesn't seem to be a relationship between ratings and budget. This is not surprising, as documentaries are generally rated higer, and horror movies lower, as compared to all other genres, even though both of them usually have a lower budget.
- However, a very high budget somewhat ensures that quality is not subpar, as all the movies with budget $>$ \$100 million are all rated above 4.
- Longer movies tend to be rated higher.

Some additional comments about the data:

- More than 90% of the movies in the dataset are English movies, however there are lots of great non-English movies out there. In particular there are many French, Korean, any Japanese movies that are world famous.
- The dataset also has a `release_date` column. However, I didn't do any analysis using it as most of the movies in the dataset were released after 2000, and there are many famous movies from 20th century that weren't in the dataset.
- In conclusion, the dataset needs to be more diverse.