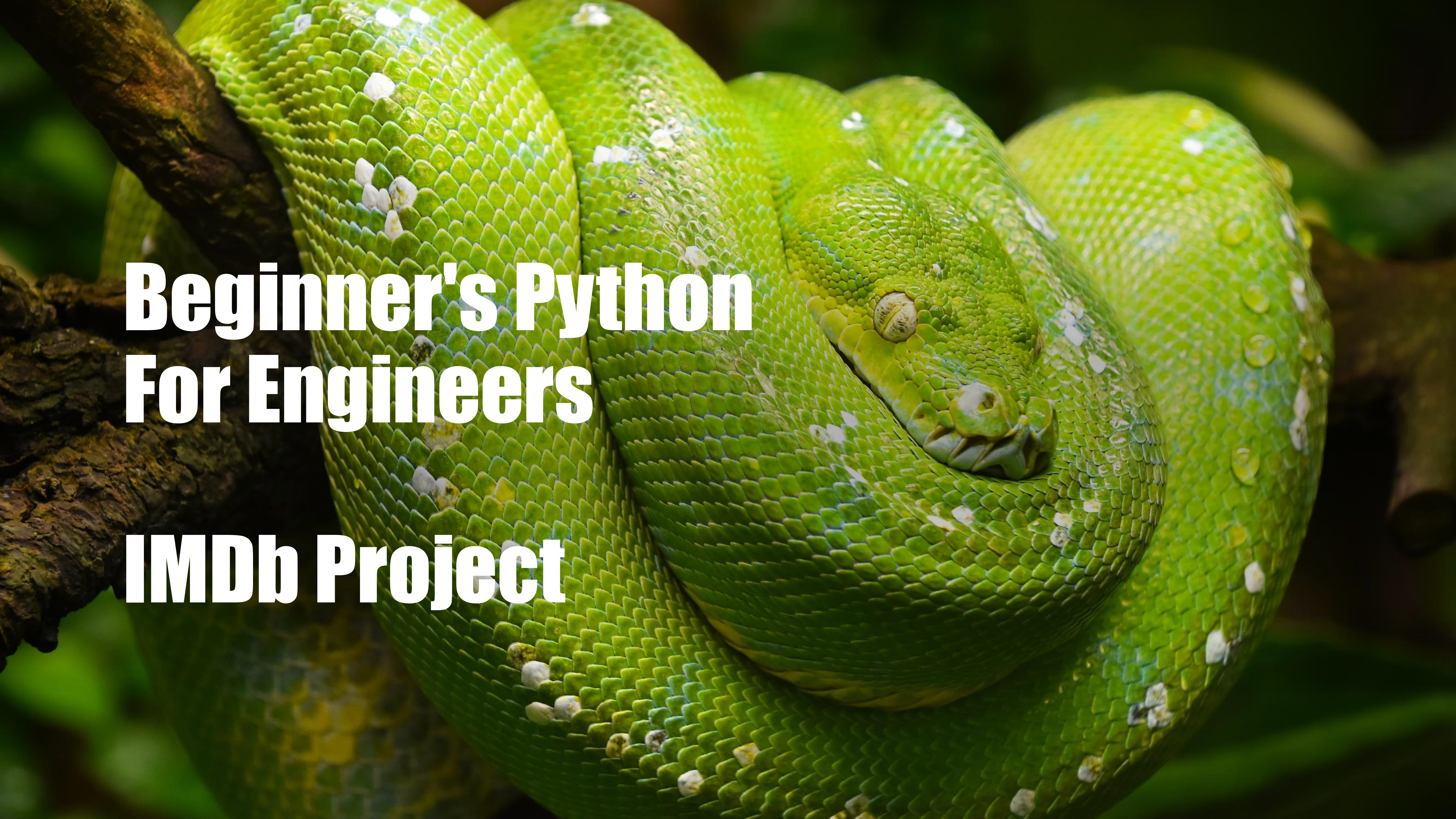
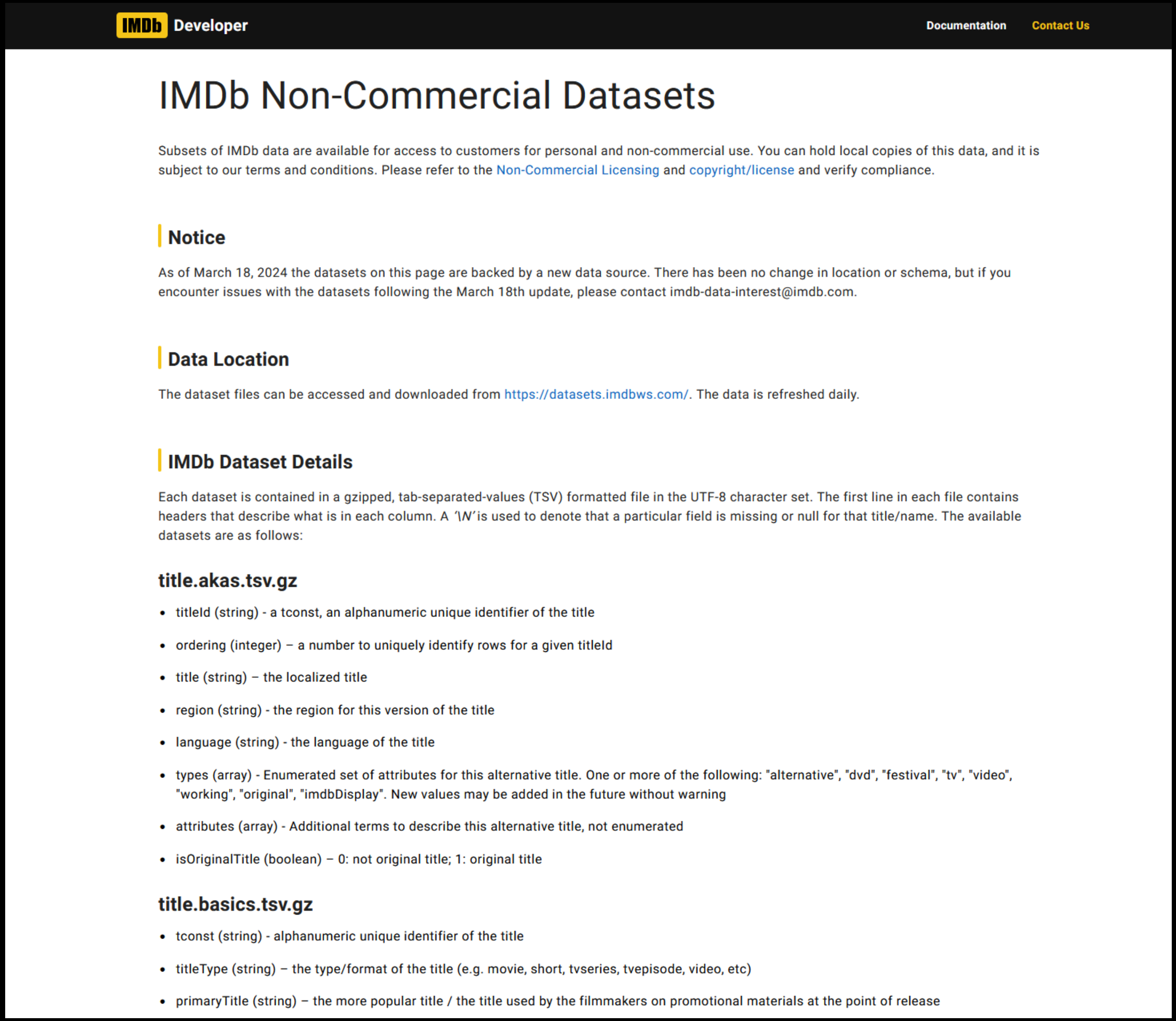


Beginner's Python For Engineers

IMDb Project

Introduction

- For this project, we used publicly available datasets from IMDb
- The base dataset is called "title basics" and contains unique identifiers of the title, release year, runtime in minutes, Adult-Boolean, title type, and genre
- The additional dataset is called "title ratings" and contains unique identifiers, average rating, and number of votes



IMDb Non-Commercial Datasets

Subsets of IMDb data are available for access to customers for personal and non-commercial use. You can hold local copies of this data, and it is subject to our terms and conditions. Please refer to the Non-Commercial Licensing and copyright/license and verify compliance.

Notice

As of March 18, 2024 the datasets on this page are backed by a new data source. There has been no change in location or schema, but if you encounter issues with the datasets following the March 18th update, please contact imdb-data-interest@imdb.com.

Data Location

The dataset files can be accessed and downloaded from https://datasets.imdbws.com/. The data is refreshed daily.

IMDb Dataset Details

Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A '\N' is used to denote that a particular field is missing or null for that title/name. The available datasets are as follows:

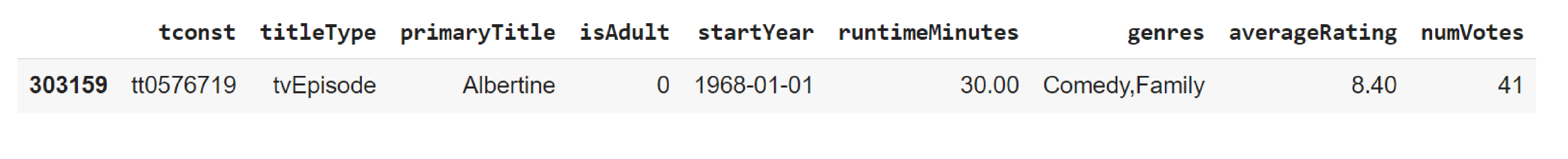
title.akas.tsv.gz

- titleId (string) - a tconst, an alphanumeric unique identifier of the title
- ordering (integer) – a number to uniquely identify rows for a given titleId
- title (string) – the localized title
- region (string) - the region for this version of the title
- language (string) - the language of the title
- types (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New values may be added in the future without warning
- attributes (array) - Additional terms to describe this alternative title, not enumerated
- isOriginalTitle (boolean) – 0: not original title; 1: original title

title.basics.tsv.gz

- tconst (string) - alphanumeric unique identifier of the title
- titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release

After Cleaning



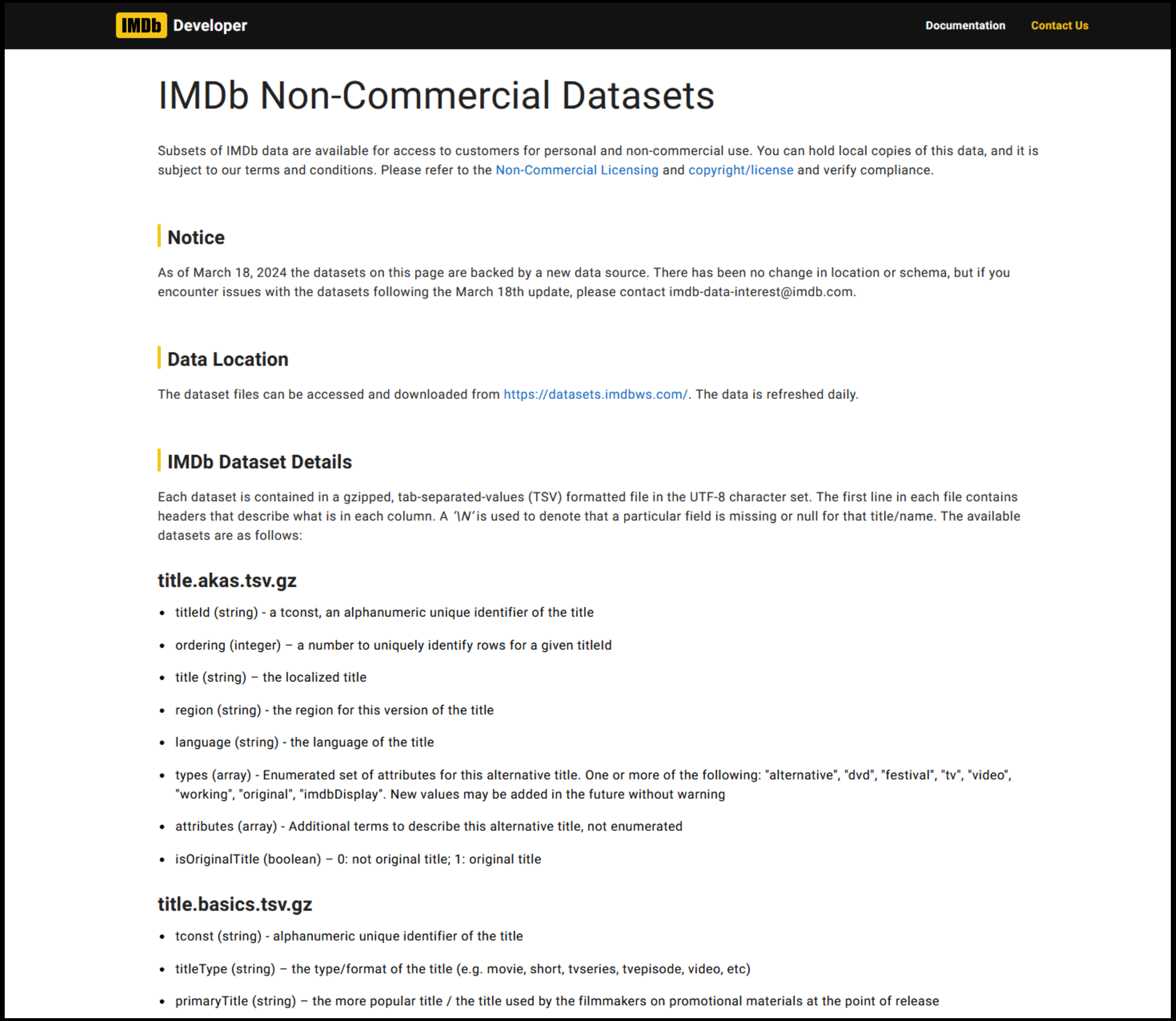
	tconst	titleType	primaryTitle	isAdult	startYear	runtimeMinutes	genres	averageRating	numVotes
303159	tt0576719	tvEpisode	Albertine	0	1968-01-01	30.00	Comedy,Family	8.40	41

Once cleaned, the data frame was comprised of

- 9 columns
- 1422330 rows

Our goals

The aim of the project was to find surprising patterns within the data, such as fluctuations of the average ratings over the years, or the distribution of percentages of adult movies over 5-year intervals



IMDb Non-Commercial Datasets

Subsets of IMDb data are available for access to customers for personal and non-commercial use. You can hold local copies of this data, and it is subject to our terms and conditions. Please refer to the Non-Commercial Licensing and copyright/license and verify compliance.

Notice

As of March 18, 2024 the datasets on this page are backed by a new data source. There has been no change in location or schema, but if you encounter issues with the datasets following the March 18th update, please contact imdb-data-interest@imdb.com.

Data Location

The dataset files can be accessed and downloaded from https://datasets.imdbws.com/. The data is refreshed daily.

IMDb Dataset Details

Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A '\N' is used to denote that a particular field is missing or null for that title/name. The available datasets are as follows:

title.akas.tsv.gz

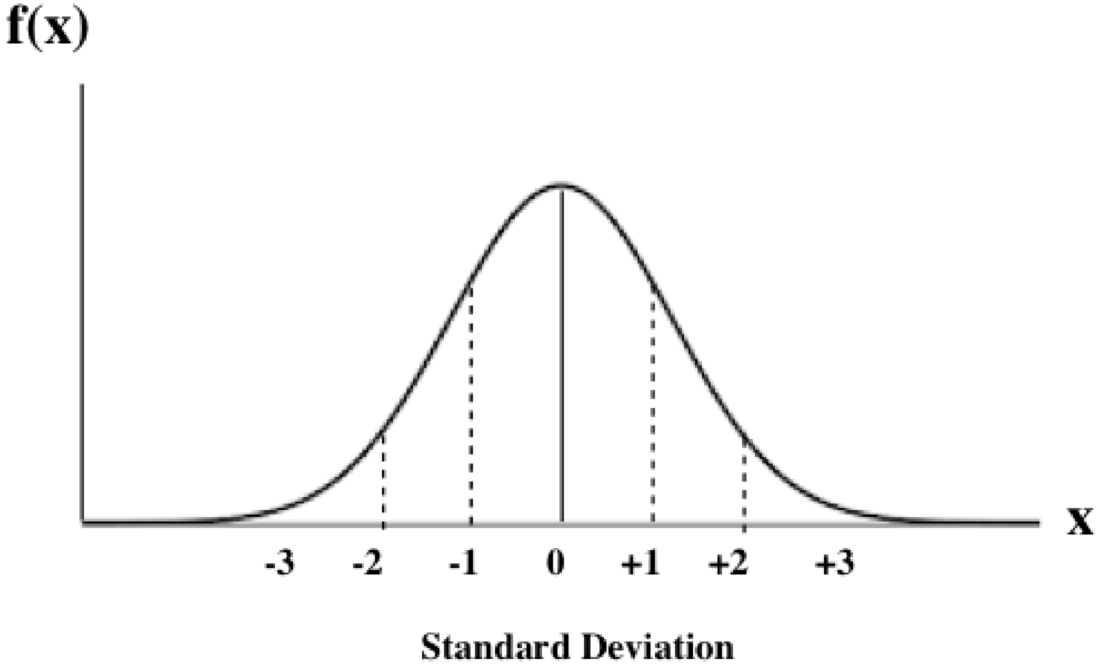
- titleId (string) - a tconst, an alphanumeric unique identifier of the title
- ordering (integer) – a number to uniquely identify rows for a given titleId
- title (string) – the localized title
- region (string) - the region for this version of the title
- language (string) - the language of the title
- types (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New values may be added in the future without warning
- attributes (array) - Additional terms to describe this alternative title, not enumerated
- isOriginalTitle (boolean) – 0: not original title; 1: original title

title.basics.tsv.gz

- tconst (string) - alphanumeric unique identifier of the title
- titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release

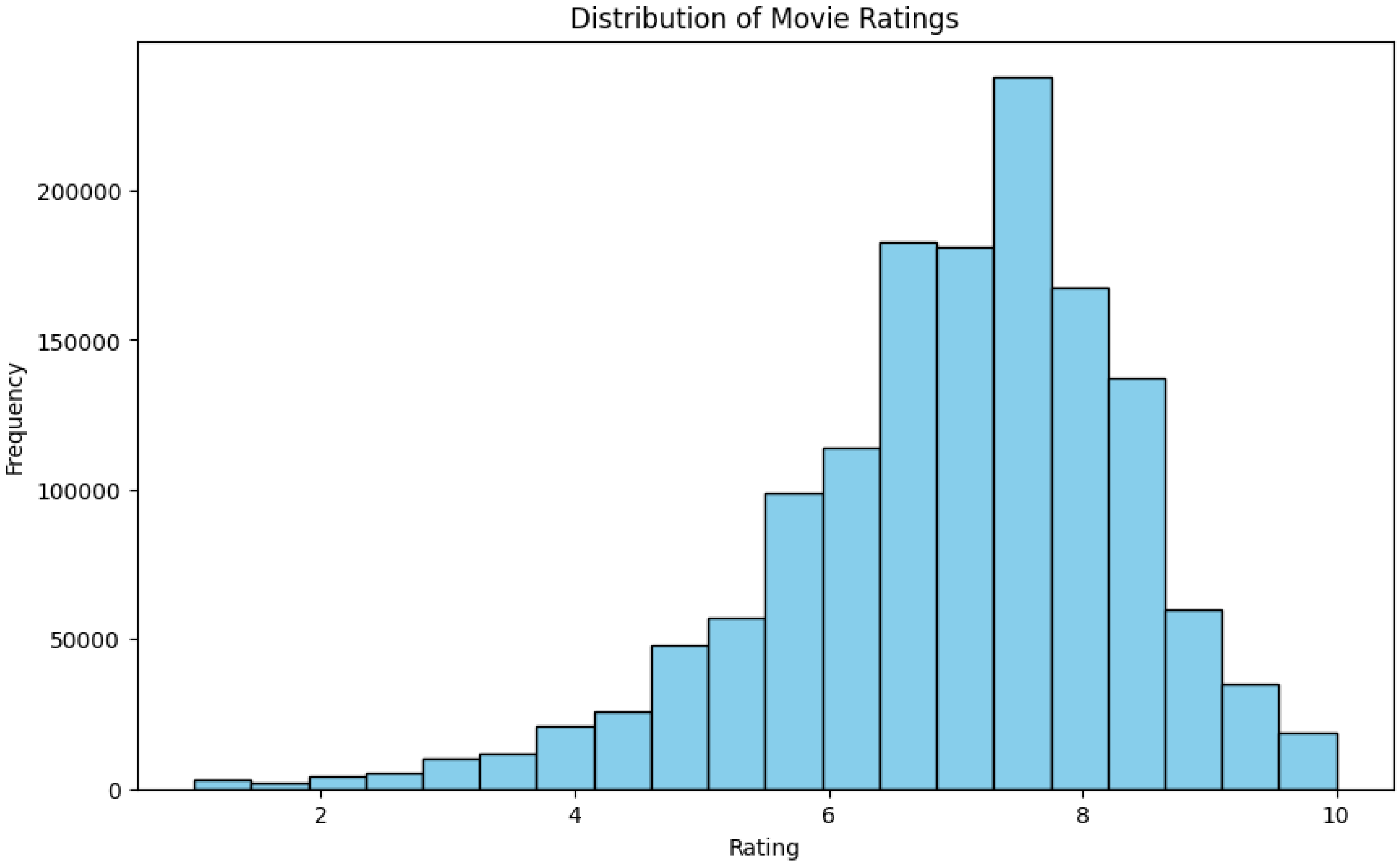
Insights and findings

What's the distribution?



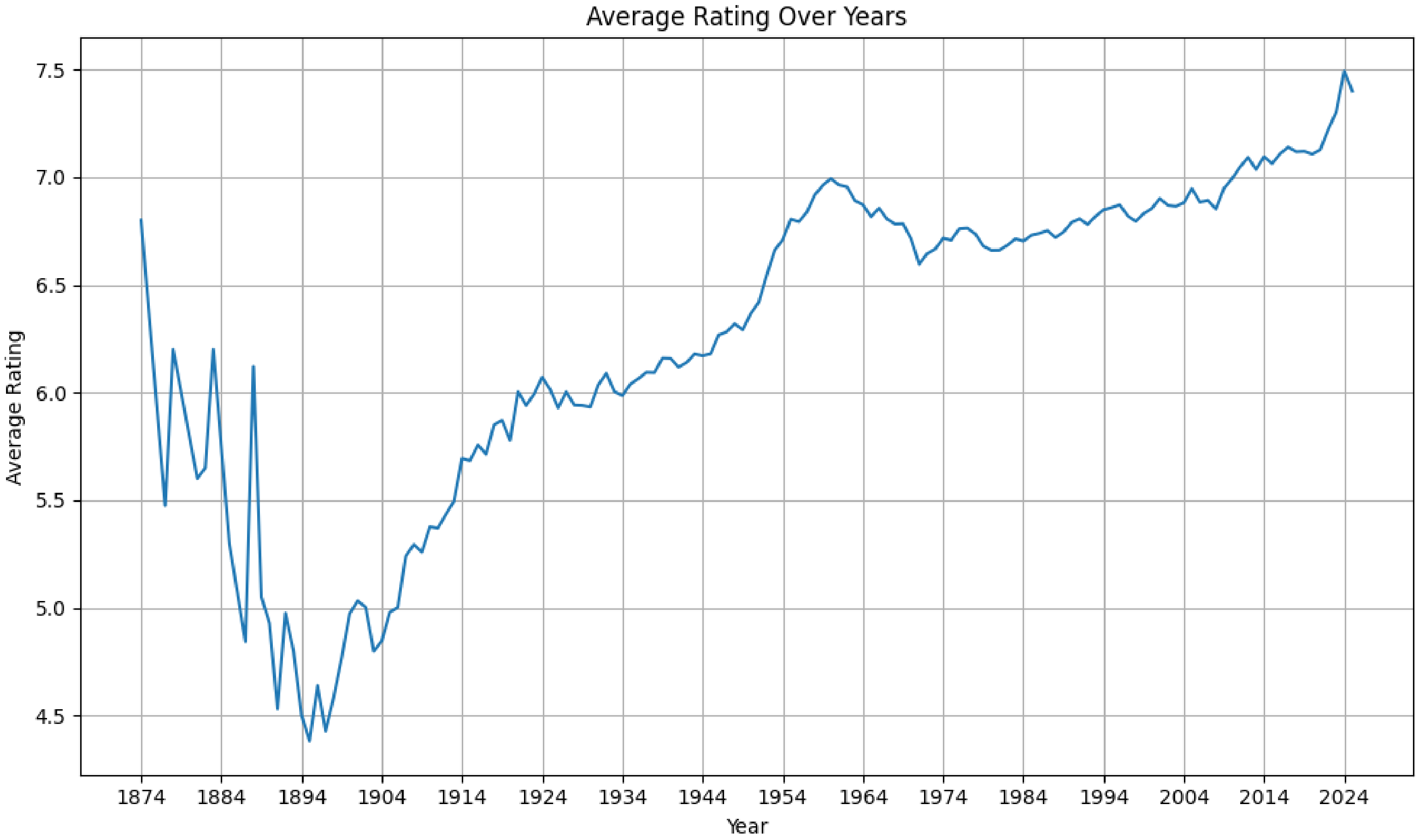
Rating Distribution

- the modal movie rating is rather high: 7.5
- people generally like movies as a form of entertainment
- people preselect movies before watching



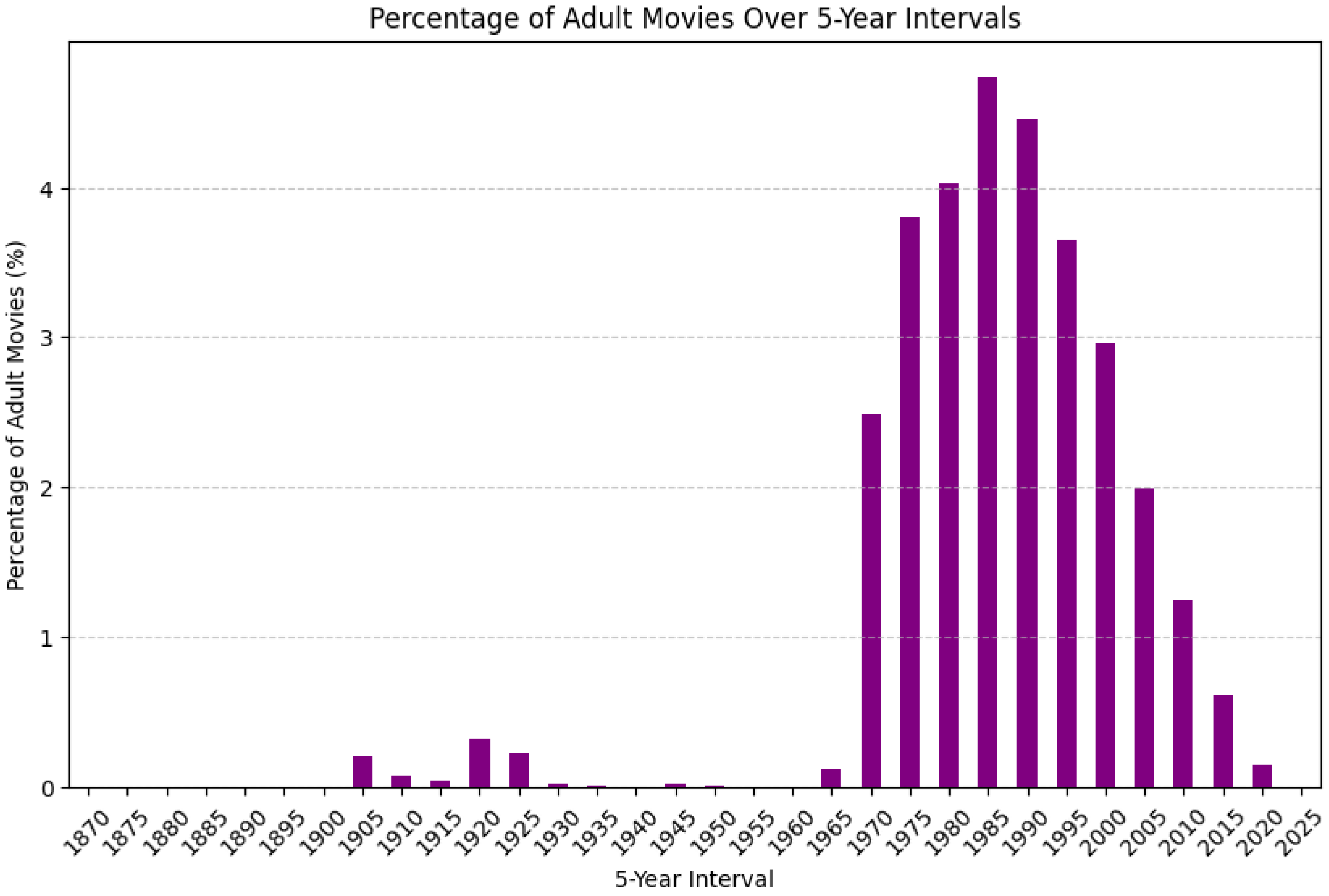
Average Rating over years

- the average ratings have been steadily growing over the last several decades
- companies are optimizing for making their titles (movies, TV-series, etc.) both more entertaining and addictive



The proportion of titles targeted specifically at adults

- peaked around 1985 but has been dropping ever since
- originally, adult movies were a taboo
- societal liberalization and the sexual revolution followed
- the revolution was followed by a reaction
- and the film studios opted out of making adult-only movies



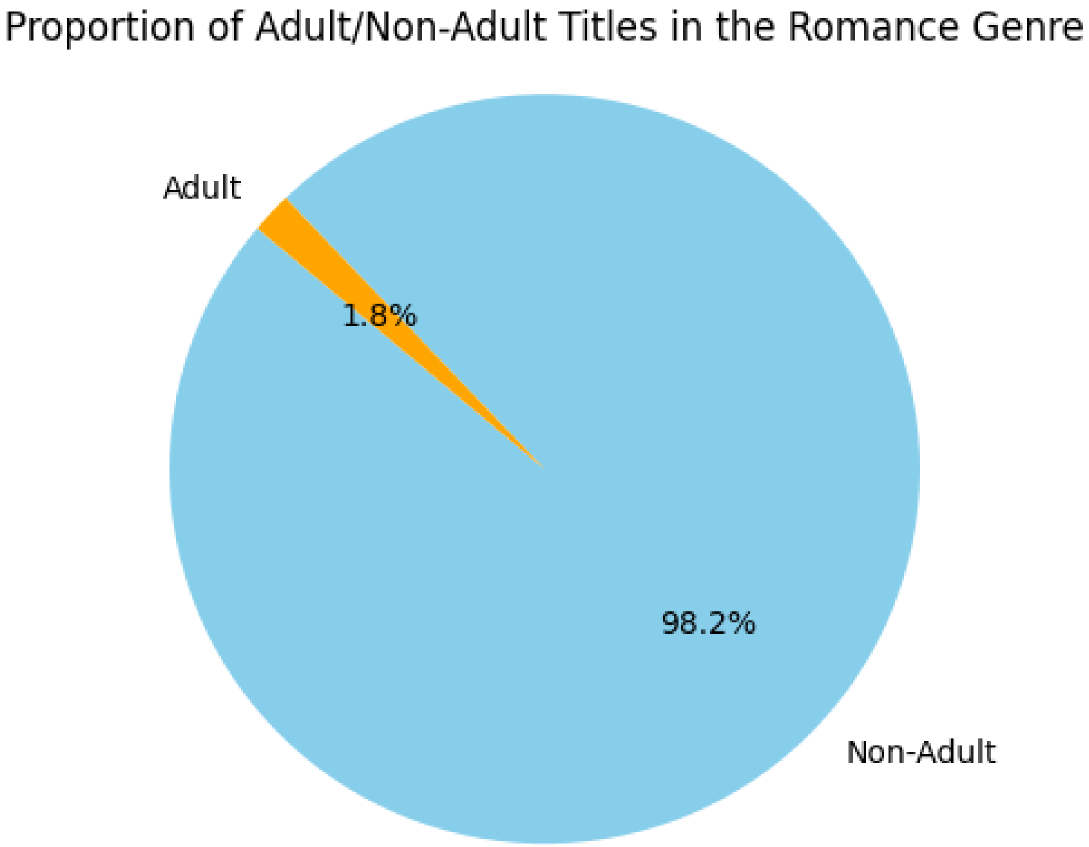
A challenge and its solution

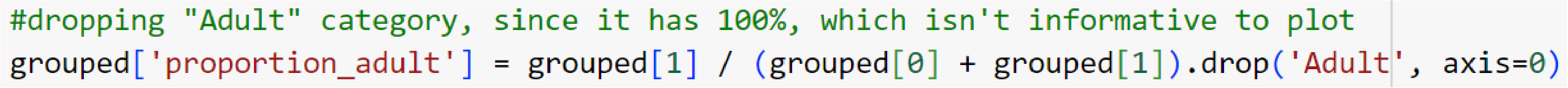
In our dataset, one intriguing challenge we encountered was addressing missing values. On one hand, fortunately enough, they are nicely denoted with a blank string instead of screwed up values. On the other hand, unfortunately, they are totally random, and as such, without proper cleaning, we can hardly extract useful data out of those columns.

Since this is a large enough dataset, a simple solution was to just replace them with null value (NaN), and pretend like they have never existed. And that was exactly what we did

A mistake &
a finding on the side

- The goal was to make a pie chart of *the genre exhibiting the highest relative* prevalence of adult content
- A whole genre named "Adult"
- Romance wins the competition!



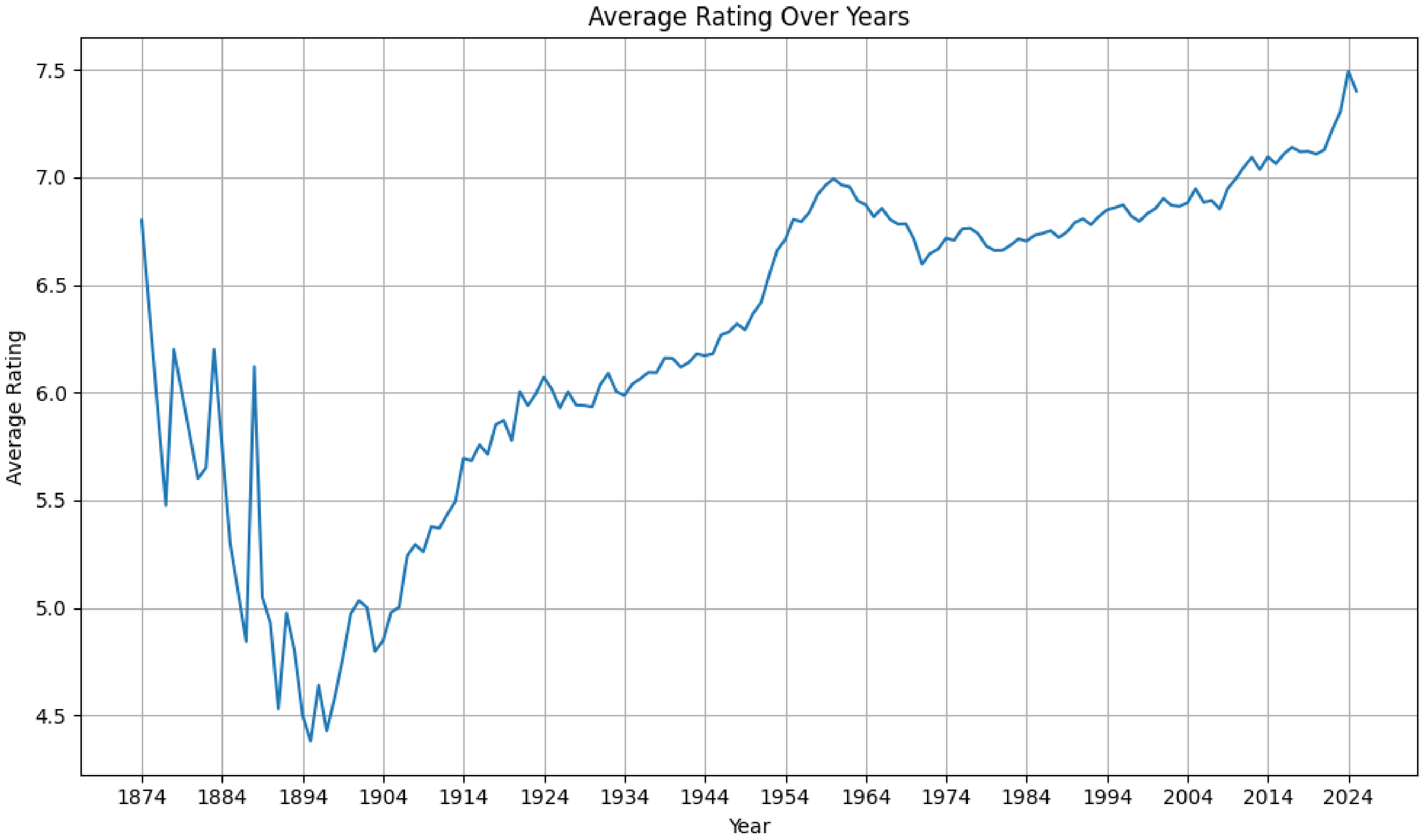


```
#dropping "Adult" category, since it has 100%, which isn't informative to plot
grouped['proportion_adult'] = grouped[1] / (grouped[0] + grouped[1]).drop('Adult', axis=0)
```

Conclusions

A description of the main finding of the project:

The average ratings have been rising for several decades



Contributions:

Anmol Kiran participated in the group project by:

- arranging meetings and participating in general group communication
- plotting three graphs significant to the nature of the data

Nguyen Duc Tung participated in the group project by:

- participating in general group communication
- completing the reading, cleaning, merging and calculation parts
- plotting a graph (Bar Plot for Highest Rating Genres)
- enhancing the layout of the project report to follow Jupyter Markdown

To Duc Minh participated in the group project by:

- participating in group communication and helped dividing group work
- doing the reading, cleaning, merging and refactoring the code

Dmitrii Gusev participated in the group project by:

- plotting two graphs (Adult/Non-Adult Titles pie-chart and Average Runtime of the Top 10 genres)
- making the presentation
- participating in the general group communication and establishing communication channels

Nguyen Vu Minh participated in the group project by:

- participating in general group communication
- plotting three graphs (Percentage of Adult Movies, Distribution of Runtime, Average Rating over Years)