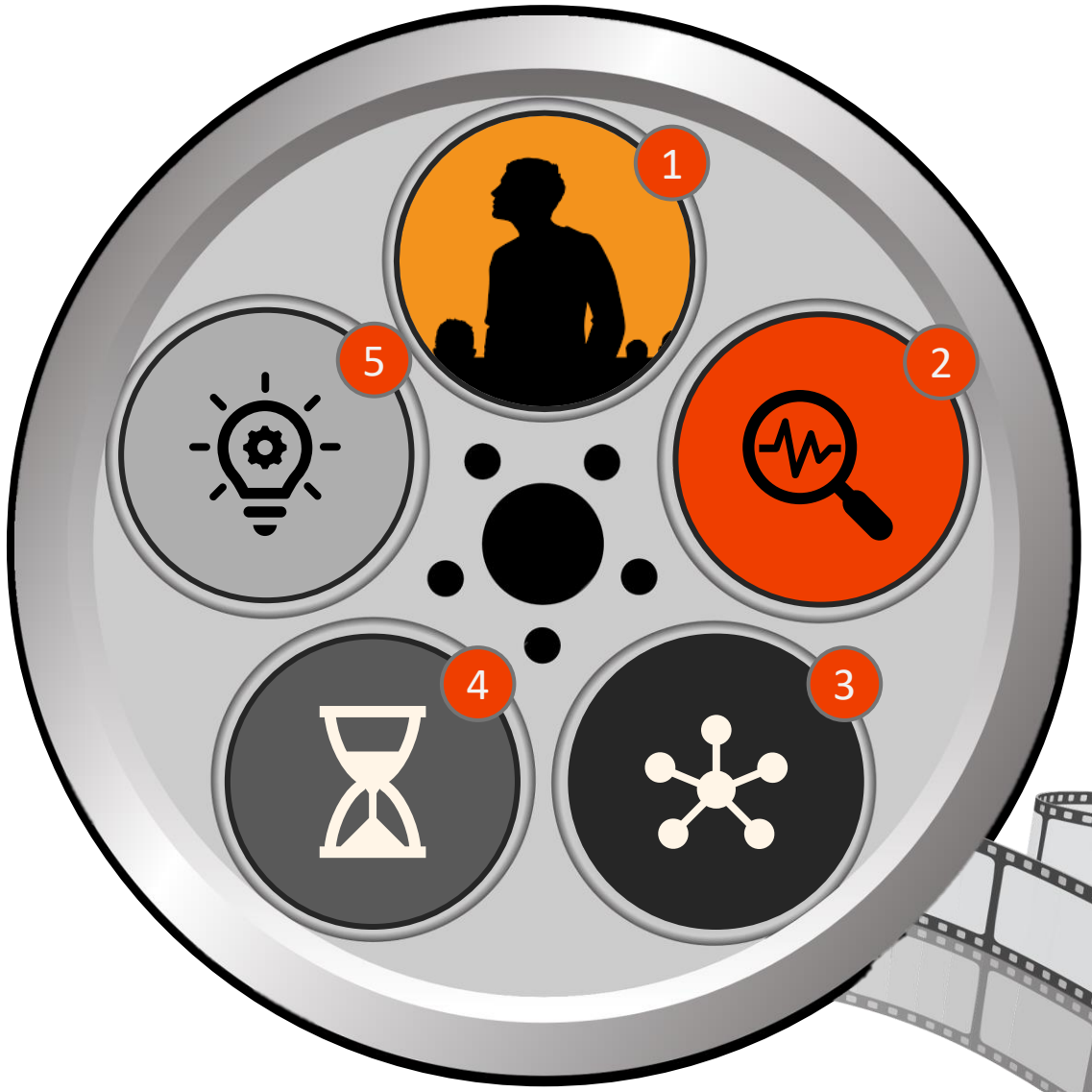

THE SECRET FOR A SUCCESSFUL MOVIE
Di Giovanni Alessio - 3091714
Du Shengqian Paolo- 3093843
Ranise Alessandro - 3219052
Roncaglia Lorenzo Sergio - 3105073

Roadmap



The Space
- Desk Analysis
- Managerial Objective
- Data Collection

Descriptive Analysis
- Data Exploration
- Data Preparation
- Univariate and Bivariate Analysis

Predictive Analysis
- Evaluation Criteria
- Models Overview
- Model Comparison

Time Series Analysis
- Exploratory Analysis
- Models & Comparison
- Conclusion & Forecast

Conclusion
- Managerial Implication
- Limitations
- Annex


CHOICE OF
BUSINESS
CASE

I. Choice of Business Case – *Cinema Industry in Italy*

Cinema Market Data | November 2023

- In November 2023, the box office **revenues** in Italy **reached 447.9 millions €**.
- Compared to the previous year, this **value increased by 65.5%** signaling a robust rebound in the industry after the Covid-19 pandemic.
- In terms of spectators, **63.2 million tickets** were sold in 2023.
- This data further underscores the sector's recovery, with a **notable increase of 58.7%** compared to the previous year.

Cinema Market Data | 2019

- While previous data seems to indicate a healthy and growing industry, we need to compare 2023 data to pre-pandemic data. In **2019**, box office **revenues** reached **€582.7 million**.
- This highlights how the **2023** data is **23% lower than 2019**.
- The number of spectators in **2023** is also **significantly lower than the 582.7 million in 2019**.
- Compared to 2019, therefore, **29.3% fewer tickets were sold**.

Fonti:
https://www.fortuneita.com/2023/11/23/tutti-segnali-di-una-stagione-entusiasmante/#:~:text=The%20Space%20%E2%80%93%20invece%2C%20con%20una,%2C%20Marche%2C%20Molise%20e%20Basilicata.
https://boxofficebiz.it/news/multiplex-in-italia-chi-sale-e-chi-scende-a-livello-di-presenze/

I. Choice of Business Case – *The Space Cinema Overview*

In 2023, **The Space Cinema** overtook UCI Cinema, becoming the **leader** in Italy by **market share**, with **19.2%** against the competitor's 18.4%. These two are the only two national cinema circuits.

In 2023 it sold more than **13 million tickets**, an **increase of 60.4%** compared to 2022 (8.1 million), beating the market.

Despite this, **17.7 million tickets** were sold in **2019**. Compared to pre-pandemic levels in 2019, **26.6% fewer tickets** were sold in 2023.

The Space has **362 cinema halls** available throughout Italy, for a total of more than **79,000 seats**.

In 2022 the **revenues** reached **75.633.254€**

Fonti:
https://www.truenumbers.it/box-office/
https://www.ufficiocamerale.it/3102/the-space-cinema-1-spa

I. Choice of Business Case – *Dataset Origin*

Data collection process

The original source of the dataset can be found at the following link:

https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies/data

The author of the dataset is themoviedb.org, The Movie Database (TMDB) is a community built movie and TV database. Every piece of data has been added by its community dating back to 2008. The TMDB is a comprehensive movie database that provides information about movies, and it is updated daily.

Furthermore, we used the same dataset to for the Time Series Analysis.

I. Choice of Business Case – *Research Goals*

We selected this dataset assuming that it could provide valuable insights that could be leveraged by a cinema chain like The Space. These insights could inform decisions regarding customer film preferences, optimizing offerings and mitigating risks, build a business model that would bring more revenues, and planning both budget and promotional offers. In particular, with this dataset we aim to:

1. Gain **consumer preferences insights** to **enhance** the company's **value proposition** and **competitive advantage** over competitors, thereby improving their **competitiveness** in the market.

2. **Optimize the offerings** by selecting only films that will be successful, thereby **reducing risk** of proposing films that will not succeed, through enhanced knowledge of consumer preferences.

3. **Develop a data-driven tool** to integrate into The Space's business model for **increasing revenue** through strategic decision-making.

4. Develop a tool to **forecast the number of films to be released and when** they are going to be released, enabling more effective **budget planning** and **customer-facing promotional strategies**.

I. Choice of Business Case – *Research Goals*

In particular, the insights provided by the results of our analysis can be leveraged to specifically target these two objectives:

Long-term Customer Retention

By consistently screening **high quality movies**, movie theatres can enhance their **brand image** by associating themselves with **quality entertainment**.
This might help them to stand-out from other competitors, and attract customers who prioritize quality.
As a long term goal, the theatre can potentially increase **customer retention** and strengthen their reputation.

Effective Scheduling

Secondary goal is to help theatres to **strategically schedule potentially movies** during peak times to maximize audience attendance.
With time series analysis we can also identify and take advantage of **seasonal trends**.


DATA
PREPARATION

II. Data Preparation – *Data Outlook*

The dataset represents a sample of 1,012,887 movies of various genres produced from the late 19th century to the present day in 2024. Overall, each observation is made up of 19 attributes (i.e., variables) that are explained below:

Name	Type	Level of measurement	Observations	Description
id	Categorical	Nominal	1,012,712 unique values	Identification number
title	Categorical	Nominal	873,026 unique values	Title of the movie
vote_average	Numerical, continuous	Ratio Scale	0 - 10	Rating's average
vote_count	Numerical, discrete	Interval	0 – 34,495	# of votes
status	Categorical	Nominal	6 unique values	Movie status
release_date	Categorical	Ordinal	1800 - 2099	Release date
revenue	Numerical, continuous	Ratio Scale	-12 $ - 3,000,000,000 $	Revenue of the movie
runtime	Numerical, continuous	Ratio Scale	- 28' – 14,400'	Duration of the movie
adult	Categorical	Nominal	2 unique values	If only adults can watch it
budget	Numerical, continuous	Ratio Scale	0 $ - 900,000,000 $	Budget of the movie
original_language	Categorical	Nominal	173 unique values	Movie's original language
overview	Categorical	Nominal	807,611 unique values	Official plot
popularity	Numerical, continuous	Ratio Scale	0 – 2,994,357	Not specified popularity index
poster_path	Categorical	Nominal	Available for 725,100 obs	Link to the poster
tagline	Categorical	Nominal	Available for 140,066 obs	Movie's catchphrase
genres	Categorical	Nominal	12,722 combinations	Combination of movie's genres
production_companies	Categorical	Nominal	194,120 combinations	Production companies involved
production_countries	Categorical	Nominal	9,822 combinations	Where the movies was made
spoken_languages	Categorical	Nominal	6.912 combinations	Languages in the movie

II. Data Preparation –*Main Issues and Dataset Limitations*

A thorough examination of the AS-IS dataset format uncovers certain characteristics of recorded variables that could pose challenges during the investigation. Therefore, additional adjustments are necessary to maintain analytical coherence, ensuring optimal readability and interpretability of the dataset.

Vote_average, vote_count and budget

1

Over 50% of values in the **vote_average and vote_count, and over 75%** for **Budget** fields are zero. We assume that in cases where this data is **N/A**, the dataset **automatically converted them to 0.** Also, **budget** shows **negative values**.

Status

2

In the **Status** column, there are six different categories. For our analysis, we are **only interested** in films that have already been **released**.

Release_date

3

The range of **release_date** spans from 1800 to 2099. The first "film" was released in 1865, and films beyond 2024 cannot have been released yet.

Revenues & budget

4

Negative revenues and negative budget have been found in the dataset, which are not interpretable. The values are expressed in **US Dollar**, but there is no information about **inflation**, so we suppose values take inflation account.

Runtime

5

The dataset contains **negative runtimes**, which cannot be interpreted. Furthermore, according to the definition, only video productions exceeding **52 minutes and last less than 240 minutes** can be considered as films.

Unnecessary variables

6

Some variables present **challenges in interpretation**, such as "popularity," while others, like poster_path or tagline, **do not contribute** to the advancement of our model.

Other

7

Variables such as **genres**, **production_companies**, **production_countries**, and **spoken_languages** are represented as combinations of values, making analysis challenging.

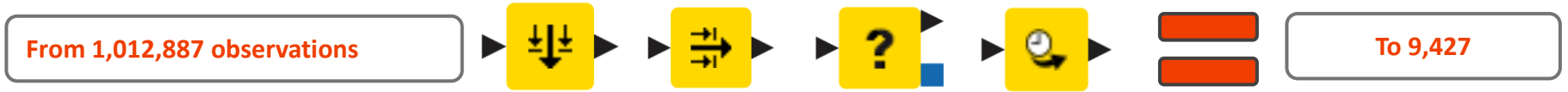
Limitations

8

The dataset **lacks** comprehensive information regarding the total **duration a movie was screened in theaters**. Furthermore, overall, the **movies have few ratings**. Even the most famous films have fewer than 35,000 votes.

II. Data Preparation – *Data Cleaning*

After analyzing the original dataset, the initial step to commence the analysis involves loading the file onto KNIME using the CSV Reader node. Once loaded and confirmed that it is displayed correctly without any alterations to the data, the data cleaning phase can be initiated. This phase involved six steps and the utilization of four distinct nodes, often in various combinations.



1. We used the *Column Filter* node to **exclude irrelevant variables**. This action helps streamline the dataset, allowing for a more expedited analysis process.

2. In the second step, a combination of the *column filter* and *row filter* nodes was employed to **maintain only the released films** and subsequently further streamline the file. Firstly, films that were not released were excluded, followed by the removal of the no longer useful variable.

3. Now, the investigation of *missing values* can be conducted using the respective node. We have opted to use **listwise deletion for each missing value** that we found, thus removing entire cases where any missing value is present.

4. Considering that some variables are encoded with a value of 0 when they are missing, the previous node must be augmented with manual removal of these values. To achieve this, we applied listwise deletion using a series of *row filter* nodes to **eliminate all values <=0** using **listwise deletion**.

5. By combining *the Extract Date & Time Fields* node with the *Row Filter* node, we ensured that **no films dated before 1895 remained**, as this marks the year when the Lumière Brothers released the first-ever motion picture. Simultaneously, we **excluded films released after 2023** to avoid any potential distortions due to recent releases, possibly affecting films from 2024. Also in this case, the data were excluded **listwise**.

6. To conclude, we opted to include productions classified as "feature films“ (*lungometraggio*) according to the definition provided by the Italian Ministry of Culture. Consequently, we **excluded** "short films“ *(cortometraggio)* that are **shorter than 52 minutes and films exceeding 240 minutes in duration**. To accomplish this, we utilized the row filter node and removed the data using listwise deletion.

II. Data Preparation – *Univariate Analysis and Understanding Data*

To enhance further understanding of the dataset and prepare it effectively for predictive modeling, we will conduct a more detailed examination of the variables through univariate analysis. This approach will enable us to comprehend variable distributions, identify outliers and significant characteristics, and refine dataset cleaning as needed. These steps aim to optimally prepare the dataset for predictive model construction.

QUANTITATIVE

	Univariate Analysis	Insights and Manipulation
Vote_average	• The variable distribution resembles that of a **normal distribution**, with a skewness level of -0.222 and a kurtosis of 2.49. (Appendix 1.1) • The values range from 0 to 10, with a mean of 6.43 and a median of 6.47. Overall, it appears that **movies are generally liked**. • There is a **peak** at maximum level.	• Given that this variable best represents the appreciation of a film, we consider it a **potential target variable.** • However, we believe **further analysis** should be conducted by comparing the data with the number of votes, as otherwise, there is a risk of distortions due to high votes and low votes.
Vote_count	• The range is **from 1 to 34,495**, with a mean of 1,620.,5 and a median of 473. (Appendix 1.2) • The distribution exhibits a skewness of 3.9 and a kurtosis of 20.3. • Executing the box plot highlights all **outliers** above 4,029, which comprise **11%** of the total. (Appendix 1.3)	• The distribution exhibits **positive skewness**. • As previously noted, many films have few votes, and those with many votes are considered **natural outliers**.

II. Data Preparation – *Univariate Analysis and Understanding Data*

To enhance further understanding of the dataset and prepare it effectively for predictive modeling, we will conduct a more detailed examination of the variables through univariate analysis. This approach will enable us to comprehend variable distributions, identify outliers and significant characteristics, and refine dataset cleaning as needed. These steps aim to optimally prepare the dataset for predictive model construction.

QUANTITATIVE

	Univariate Analysis	Insights and Manipulation
Revenue	• The values **range from 1 to nearly 3,000 M**, with a **mean** of **70.3 M** and a **standard deviation** of **160 M**. (Appendix 2.1) • It exhibits a **kurtosis** level greater than 50 and a **skewness** of 5.7. • A large number of **outliers**, corresponding to **12%** of the total, have been identified. (Appendix 2.2)	• **Successful movies can achieve very high** revenues, so we assume that the **outliers** are **natural**. • Revenues could be the **potential target** of our model. In that case we will have to perform some **transformations** to make up for the peakness and skewness of the distribution. **Further considerations** will be made in the section of target variable selection
Runtime	• The **average length** of a movie is **109.5 minutes**, with a standard deviation of 21.89. (Appendix 2.3) • The distribution exhibits **positive skewness,** with some films nearly reaching four hours in length. • **Outliers** have been identified, comprising **3.5%** of the total.	• The objective of analyzing a movie's runtime is to identify the **optimal duration** for a movie, i.e., the duration that allows the movie to receive favorable ratings. • In this case as well, we believe that the present **outliers** are **natural**, so no further manipulations are necessary.

II. Data Preparation – *Univariate Analysis and Understanding Data*

To enhance further understanding of the dataset and prepare it effectively for predictive modeling, we will conduct a more detailed examination of the variables through univariate analysis. This approach will enable us to comprehend variable distributions, identify outliers and significant characteristics, and refine dataset cleaning as needed. These steps aim to optimally prepare the dataset for predictive model construction.

QUANTITATIVE

Budget

Univariate Analysis	Insights and Manipulation
• The budget has a **maximum** value of **460 M**, a **mean** of almost **25 M**, and a **standard deviation** of **38.7 M**. (Appendix 3.1) • The distribution exhibits a **skewness** of **3.23** and a **kurtosis** of **14.3**. • **Outliers**, comprising **9.6%** of the total, have been identified in this case. (Appendix 3.2)	• A **high standard deviation** is due to the large range of values that the variable can assume. The **range for budgeting a movie is therefore very wide.** • To enhance the distribution's adaptability to models, we propose to build a **dummy** to detect the **high budget** movies. • We assume that the **outliers** are **natural**.

II. Data Preparation – *Univariate Analysis and Understanding Data*

To enhance further understanding of the dataset and prepare it effectively for predictive modeling, we will conduct a more detailed examination of the variables through univariate analysis. This approach will enable us to comprehend variable distributions, identify outliers and significant characteristics, and refine dataset cleaning as needed. These steps aim to optimally prepare the dataset for predictive model construction.

QUALITATIVE

	Univariate Analysis	Insights and Manipulation
Original language	• The dataset display **59** different original language. (Appendix 4.1) • **English** represents **78.47%** of the values. • The **second** most common language for films is **Hindi** with **3%.** • **Italian** films account for **0.987%** of the total.	• **English** is the most represented language in film productions. • As expected, the second most common language is **Hindi**, from India, where there is a strong film production industry. • Due to the fragmentation of the variable, **further manipulation** will be required to manage this granularity.
Others variables	• As highlighted earlier, many qualitative variables are represented as **combinations of values**. Therefore, it is **not currently possible to conduct univariate analysis**. • The study of the **release date** will be addressed in the **time series analysis section**.	• **Manipulations will be conducted** to enable a more in-depth study of variables such as genres or production companies. Therefore, further details will be provided in the following slides.

II. Data Preparation – *Building Target Variable*

As specified in the introductory section, our main goal is to build a model capable of predicting potential successful movies. A successful film can be defined by two components: qualitative (content) and quantitative (revenues). By offering successful movies, we achieve two benefits.

- **Demand side:** we provide viewers with high-quality content, reducing the likelihood of them leaving theaters disappointed.
- **The Space's side:** proposing successful films not only satisfies our customer base but also attracts more spectators and enhance the revenues profits (film revenues are correlated by cinema revenues, since for each ticket sold about 50% of the sales price is paid to the film's producer).

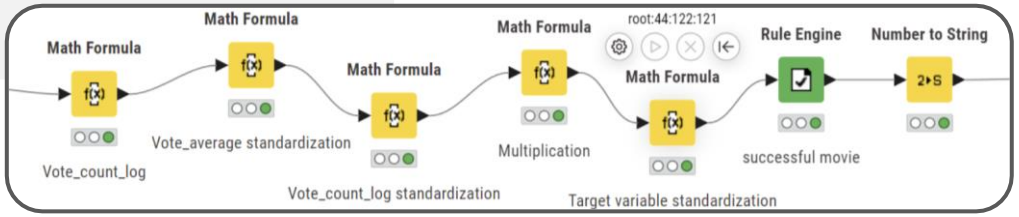
To measure a film's success, we conducted a series of computational steps using Math Formula Nodes:

1. We **logged** the variable 'vote count' due to its extremely skewed distribution and **normalized** it within the range of 0-1.
2. Similarly, we **normalized** the 'vote average' variable.
3. Then we **combined** the vote average and vote count to capture both the perceived quality and popularity/reach of the movie, resulting in a more robust indicator for decision-making.
    1. In this way we were able to obtain a variable that assigns a score from 0 to 1 to each movie based on the **equally weighted vote average and vote count**. As a result, we obtained a metric that assesses both the quantitative and qualitative sides of success for a movie.
    2. Click here for vote count logged and weighted rating standardized distribution
4. Therefore, using the **Rule Engine** node, we created our ultimate target variable: 'successful_movie,' a dummy that indicates whether a movie is successful if it exceeds a threshold of **0.6** in the 'weighted_rating' variable.
5. Finally, through the **Number to String** node we transformed the target variable into a string (for computational reasons in order to run the predictive models correctly)

Our choice can easily be explained with an example: a film with one vote count and a vote average of 10 has not the same weight of a film with a vote count of 1.000 and a vote average of 8.

We did not also consider revenue as a target variable for two main reasons.

- **"Vote count"** and **"revenue"** are **highly correlated** (following slide)
- Through sample testing, we noticed that **"revenue"** does not always correspond to the **truth** as opposed to vote average and vote count which are exact metrics as they are extracted directly from TMDB



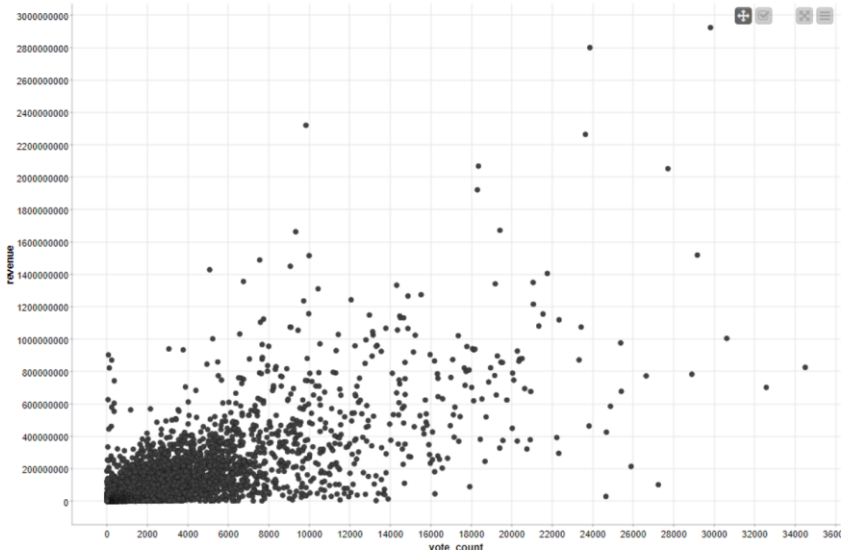
Lash, Michael T., and Kang Zhao. "Early Predictions of Movie Success: The Who, What, and When of Profitability." *Journal of Management Information Systems*, vol. 33, no. 3, 2 July 2016, pp. 874–903, www.biz.uiowa.edu/faculty/kangzhao/pub/JMIS_2016.pdf, https://doi.org/10.1080/07421222.2016.1243969.

II. Data Preparation – *Building Target Variable: Spoiler*

Why not predicting revenues?

Another reason:

First column name *String*	Second column name *String*	Correlation value *Number (double)*	p value *Number (double)*	Degrees of freedom *Number (integer)*
vote_count	revenue	0.748	0	9425



Another reason we excluded revenues from the target variable is its **strong correlation** with the vote_count variable. By doing so, we avoided incorporating two correlated variables into the target variable, which could lead to **distortions**.

Furthermore, we chose vote_count over revenues for the reason mentioned earlier: after double-checking with real data, we discovered that some values of the revenue variable did not correspond to the truth. This could lead to **significant distortions** in the predictive model and render our model **ineffective**.

II. Data Preparation – *Constructing Variables for Bivariate Analysis*

Production Company

This variable shows the name/s of the production company/s involved in the production of the movie. Since **more than one company can be shown for a single observation**, we need to apply some transformations in order to make the data more clear and actionable for the analysis.

By applying a sequence of nodes, we isolated the single production companies (9796 unique companies) and created the following variables:

Production Company Dummies

We created a dummy variable for the most famous companies with more than 100 observations and grouped the rest into a new variable called “Minor_Company”, resulting in 21 new dummy variables.
These dummy variables will be useful for the logistic regression model.

Number of Prod_Companies

We also created a variable that specifies, for each movie, the number of production companies that were involved in the production of the movie. This will hopefully give us some additional insights.

Main_Production_Company

Since we noticed that the production companies were shown ordered by the one which gave the major contribution in the production of the movie, we isolated the first company of the list and created the new variable “Main_Production_Company”. This variable will be useful to have an easier visualization of the descriptive statistics.

Main_Company_Size

To further simplify the visualization of the bivariate analysis, we grouped the production companies into three classes {small, medium, big}, where we consider small companies the ones having less than 10 observations, medium companies between 10 and 100 observations and big companies more than 100 observations.
ATTENTION: this variable doesn’t represent a classification based on the real number of movies produced by the production companies, but the creation of this variable will be useful to have a clearer visualization of the statistics.

II. Data Preparation – *Constructing Variables for Bivariate Analysis*

We encountered the same issue for the variables "*Genre*", "*Production Country".* With the same logic applied for the variable "*Production Company*", we isolated the single values and applied the following transformations.

Genre

Genre Dummies

We have 19 unique genres in total, with Drama and Comedy being the most observed, while TV movies and Documentaries being the least observed.

Number of Genres

Indicates the number of genres a movie is classified into. This variable could give us additional information on genre synergies.

Main Genre

Indicates the main genre a movie is classified as.

Production Country

Production Country Dummies

We have 119 unique production country in total, with United Stated and United Kingdom being the most observed.

Number of Countries

Indicates the number of country where a movie is filmed.

Main Country

Indicates the main country where a movie is filmed.

II. Data Preparation – *Constructing Variables for Bivariate Analysis*

To conclude the data manipulation section, we can proceed to analyze in more detail the "*original language*" and the "*Budget*" of the movie.

Budget

Original language

High Budget Dummy

Genre Dummies

We have constructed the variable to determine whether a movie has a high budget or not. According to the Treccani encyclopedia, the level at which a budget can be considered high varies depending on the country of production, but a good approximation for a high budget globally can be around 30 million. Therefore, we have set this threshold to study whether a film can be considered high budget or not.

We have a total of 59 unique original languages. For the manipulation, we decided to retain the original languages that are used in more than 90 movies. The original languages spoken in fewer than 90 movies have been grouped into the category "Minor_language".

Fonti:
https://www.treccani.it/enciclopedia/economia-del-cinema_(Enciclopedia-Italiana)/

II. Data Preparation – *Univariate Analysis After Data Manipulation*

To enhance further understanding of the dataset and prepare it effectively for predictive modeling, we will conduct a more detailed examination of the variables through univariate analysis. This approach will enable us to comprehend variable distributions, identify outliers and significant characteristics, and refine dataset cleaning as needed. These steps aim to optimally prepare the dataset for predictive model construction.

QUANTITATIVE QUALITATIVE

	Univariate Analysis	Insights and Manipulation
Budget	• After manipulation, we can observe how film productions are divided based on the budget. (Appendix 5.1) • From the data, we can see that **25.78%** (2430) of movies are produced with a high budget. • Conversely, **74.22%** (6997) of movies are produced with a low budget.	**25%** • As expected, the number of films produced with a high budget is significantly lower compared to those with a low budget. • We can assert that in this dataset, **high-budget Movies** correspond to approximately **25% of the total films**.
Original Language	• As highlighted earlier, even with categorization, **English** remains the most represented language. • The "**Minor Language**" at this point are the second most represented category, with **811** cases, accounting for **8.6%** of the total. (Appendix 5.2)	• **English** continues to dominate the film production, surpassing **75%** of the total observations. • We will also pay close attention to films in the **Italian language**, as The Space is an Italian entity, and it would be interesting to better understand how Italian productions perform.

II. Data Preparation – *Univariate Analysis After Data Manipulation*

To enhance further understanding of the dataset and prepare it effectively for predictive modeling, we will conduct a more detailed examination of the variables through univariate analysis. This approach will enable us to comprehend variable distributions, identify outliers and significant characteristics, and refine dataset cleaning as needed. These steps aim to optimally prepare the dataset for predictive model construction.

QUALITATIVE

	Univariate Analysis	Insights and Manipulation
Production Company	• The dataset display **9797** different production company. (Appendix 6.1) • The most represented company are **Warner Bros Pictures, Universal Pictures and Paramount**. • **Warner Bros Pictures**, the most represented value, weight the **6,5%** of the total.	• The dataset exhibits **fragmented values** for production companies. • Specifically, **6646** production companies have **produced only one movie**. • After the manipulation, the **companies that produced more than 100 movies are 19**.
Number of production company	• The **majority** of movies (2309) are **produced by a single production company**. • Movies produced by **more than 3** production companies amount to **3086**, meanwhile, ones that are produced by less than 4 are **6341**.	• Generally, **fewer production companies are preferred for the production of a movie**. • However, there are **cases** where as many as **30 different companies** have worked on the same film.

II. Data Preparation – *Univariate Analysis After Data Manipulation*

To enhance further understanding of the dataset and prepare it effectively for predictive modeling, we will conduct a more detailed examination of the variables through univariate analysis. This approach will enable us to comprehend variable distributions, identify outliers and significant characteristics, and refine dataset cleaning as needed. These steps aim to optimally prepare the dataset for predictive model construction.

QUALITATIVE

Univariate Analysis	Insights and Manipulation
Main Production Company • If we consider only the **main production companies**, there are only **4022**. • In this case, the companies that have produced the most films are **Paramount**, **Universal Pictures**, and **Columbia Pictures**. • **Paramount** is the most represented, being the main production company for **3.18%** of the movies.	• The differences compared to the main production companies are a first indication of how production **companies participate in the production of many movies but are not always the primary producers.**
Main company production size • After manipulation and categorization into small, medium, and large categories, we can observe that **62.3%** of companies are **small**, **21.3%** are **medium**, and **16.4%** are **large**. (Appendix 7.1)	• Large production companies are few in comparison to the total, despite producing the majority of films.

II. Data Preparation – *Univariate Analysis After Data Manipulation*

To enhance further understanding of the dataset and prepare it effectively for predictive modeling, we will conduct a more detailed examination of the variables through univariate analysis. This approach will enable us to comprehend variable distributions, identify outliers and significant characteristics, and refine dataset cleaning as needed. These steps aim to optimally prepare the dataset for predictive model construction.

QUALITATIVE

	Univariate Analysis	Insights and Manipulation
Main Genre	• As said before, we can highlight **19 uniques** main genres. (Appendix 8.1) • The genre most represented is **Drama**, followed by **Comedy**. • Most of the movies can be descripted with less than four genres. (Appendix 8.2)	• Drama and Comedy are very popular. • Usually, a movies has not a single genre, but it is composed of a **mix of genres,** but usually less than four.
Production Country	• As previously highlighted, the nation producing the most films is the **United States**, followed by the **United Kingdom**. (Appendix 8.3) • The **United States** holds almost a **monopoly** on film production, with a share exceeding 50%. • The "**Other**" category is the second most represented.	• In this dataset, Hollywood, the **United States**, remains the top film producer. • **Italy** closes this series as the tail end, with just **113** observations.

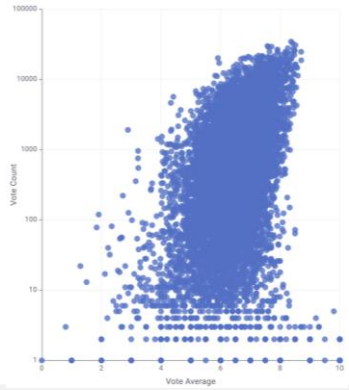
II. Data Preparation – *Bivariate Analysis*

First we wanted to uncover some patterns between our target variable and the explanatory variables. Since our target variable "*successful*" (no/yes) considers vote average and vote count, we are going to analyze them individually.

Vote Average

Vote average does not show significant correlation with the revenue and most of the explanatory variables. Most interesting correlations are:

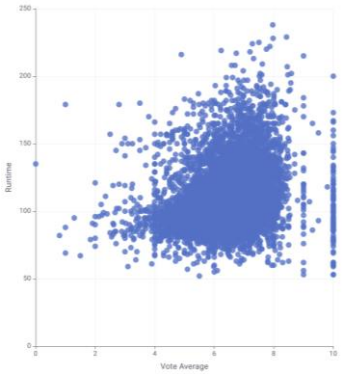
Vote Average – Vote Count



We noticed a positive correlation between Vote Average and Vote Count. But the correlation value is not so strong (correlation value of 0.297).

Since p-value = 0, we further inspected the relationship by observing the scatter-plot distribution.

Vote Average - Runtime



The significant correlation coefficient of **0.254** suggests a positive, although not so strong, association between **vote average** and **runtime**, meaning that longer movies tend to have higher ratings.

Vote Average – Main Production Country

	Group	N	Mean	Std. Deviation	Std. Error
vote_average	**Japan**	167	**6,9285**	0,9822	0,076
vote_average	**UK**	583	**6,6439**	0,899	0,0372
vote_average	**China**	148	**6,6206**	0,8393	0,069
vote_average	other	933	6,5538	1,173	0,0384
vote_average	Belgium	119	6,5503	0,8087	0,0741
vote_average	France	463	6,5415	0,8059	0,0375
vote_average	USA	5192	6,4209	0,9495	0,0132
vote_average	Australia	163	6,4197	1,0103	0,0791
vote_average	Germany	277	6,4121	0,9435	0,0567
vote_average	India	570	6,3153	1,1865	0,0497
vote_average	Italy	113	6,2217	1,2861	0,121
vote_average	Spain	114	6,2107	1,0251	0,096
vote_average	Canada	400	6,1904	0,9345	0,0467
vote_average	**Russia**	185	**5,6186**	1,4259	0,1048
vote_average	Total	9427	6,4304	1,0077	0,0104

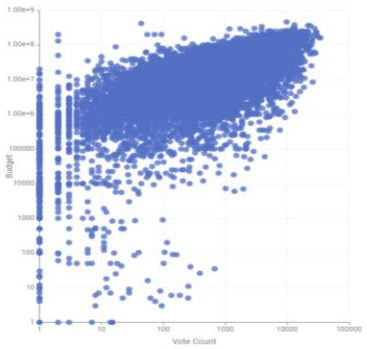
We wanted to see if movies produced in certain countries get higher average rating than others.

We noticed that, on average, movies that are mainly produced in **Japan**, **UK** and **China** get the highest average rating. On the other hand, movie produced mainly in **Russia** get the lowest rating. The F-test shows a significant value of 20.1, meaning there's a difference between group means compared to the variation within groups.

II. Data Preparation – *Bivariate Analysis*

Vote Count

Vote Count – Budget



There is a **positive correlation** between the variable **vote_count** and the **budget.** Above we can see the scatter plot with the log transformation of the vote_count on the x axis and the log transformation of the budget on the y axis. The **correlation value** is **0.587,** which confirms a very strong association between these two fundamental variables of our analysis.

Vote Count – Main Company Size

	Group	N	Mean	Std. Deviation	Std. Error
vote_count	**medium**	**2008**	**2544,0339**	**4129,6197**	**92,157**
vote_count	big	1547	2334,7757	3684,7929	93,6845
vote_count	**small**	**5872**	**1116,53**	**2188,0187**	**28,5534**

We can notice a **significant difference** in the **mean values** corresponding to **different sizes of the main company** producing the movie. Surprisingly, the ANOVA exhibits that if the main company is a **medium sized** one, the **average number of votes** for the movie will be higher. As we expected, if it is a **small one**, then the **mean vote count** will be **lower**. The high value of the F-test (225.821) confirms the significancy of the bivariate analysis, and also the Levene-Test shows the reliability of this consideration.

Vote Count – Main Genre

	Group	N	Mean	Std. Deviation	Std. Error
vote_count	Adventure	557	3.215,28	4905,088	207,8353
vote_count	Science Fiction	223	3.134,96	4316,8726	289,0792
vote_count	Animation	280	2.946,30	4043,24	241,6298
vote_count	Fantasy	205	2.834,29	3727,5119	260,3407
vote_count	Action	1451	2.191,63	3942,1332	103,4898
vote_count	Family	191	1.879,94	3102,4229	224,4834
vote_count	Horror	621	1.558,48	2105,7483	84,5008
vote_count	Mystery	120	1.549,07	2912,6324	265,8857
vote_count	Crime	440	1.509,81	2829,9727	134,9137
vote_count	Thriller	417	1.434,09	2355,7873	115,3634
vote_count	Drama	2181	1.235,49	2539,4159	54,3758
vote_count	War	93	1.221,67	2697,3726	279,7045
vote_count	Western	68	1.134,21	2354,4738	285,5219
vote_count	Romance	276	1.070,14	2120,7943	127,6568
vote_count	Comedy	2045	1.025,68	1809,3853	40,0115
vote_count	History	62	927,42	2259,1726	286,9152
vote_count	Music	96	755,67	2072,222	211,4953
vote_count	Documentary	98	172,92	308,5272	31,166
vote_count	TV Movie	3	66,33	111,4331	64,3359

The ANOVA shows that the **vote_count** variable presents **different mean values** according to the **different genres.** The F-test is significant with a value of 29.779. The higher average number of votes is presented for the genres **“Adventure”, “Science Fiction”** and **“Animation”,** while the **lower** ones concern the genres **“TV Movie”** and **“Documentary”.** As a matter of fact, the first movies reached mean vote counts around 3 millions, while the last ones are lower than 200.000. (Levene-test = 53.82)

II. Data Preparation – *Bivariate Analysis*

Revenue

Revenue – Main genre

	Group	N	Mean	Std. Deviation	Std. Error
revenue	**Animation**	**280**	**179.000.000,00**	**253000000**	**15106599,53**
revenue	**Adventure**	**557**	**174.000.000,00**	**303000000**	**12858232,18**
revenue	**Fantasy**	**205**	**134.000.000,00**	**213000000**	**14904507,08**
revenue	Science Fictior	223	133.000.000,00	261000000	17448961,05
revenue	Family	191	132.000.000,00	233000000	16892638,28
revenue	Action	1451	108.000.000,00	219000000	5748900,791
revenue	War	93	68.884.097,43	151000000	15678491,02
revenue	Mystery	120	50.294.279,59	94730800,87	8647699,421
revenue	Thriller	417	48.509.921,91	87709864,05	4295170,522
revenue	Comedy	2045	46.479.295,14	82419950,85	1822576,201
revenue	Horror	621	44.090.138,12	70963094,43	2847650,887
revenue	Music	96	42.668.309,89	111000000	11317624,79
revenue	Crime	440	40.506.981,63	86623618,79	4129618,993
revenue	Romance	276	39.541.933,59	75760039,2	4560218,474
revenue	History	62	38.212.416,13	77035832,27	9783560,482
revenue	Drama	2181	37.876.476,10	91263993,41	1954211,818
revenue	Western	68	28.443.338,13	69688736,42	8451000,623
revenue	**TV Movie**	**3**	**14.377.704,33**	**23926159,08**	**13813774,39**
revenue	**Documentary**	**98**	**11.480.249,56**	**30484878,64**	**3079437,773**

From the bivariate analyses of the revenues, the **first strong association** that we can see is between the **mean level of income** and **the main genre** of the movie. In fact, if the main genre is **"Animation", "Adventure"** or **"Fantasy",** the movie will reach **higher revenues**. On the other hand, if the movie is a **"TV movie"** or a **"Documentary",** then it will achieve lower levels of income. The F-test confirms these statements, showing a value of 45.347.

Revenue – Main Production Country

	Group	N	Mean	Std. Deviation	Std. Error
revenue	**China**	**148**	**135.000.000,00**	**196000000**	**16115904,11**
revenue	**United Kingdom**	**583**	**98.572.118,16**	**201000000**	**8309966,957**
revenue	United States of America	5192	87.216.564,68	185000000	2561280,067
revenue	Germany	277	63.839.453,68	98.406.653,20	5.912.682,75
revenue	Japan	167	62.845.812,82	150.000.000,00	11.569.238,24
revenue	Canada	400	62.688.159,40	119.000.000,00	5.968.590,54
revenue	Australia	163	59.421.878,36	99.408.029,73	7.786.237,81
revenue	France	463	41.531.958,93	75.009.192,36	3.485.972,80
revenue	other	933	33.149.658,14	101.000.000,00	3.310.414,72
revenue	Italy	113	24.734.809,45	60.011.850,38	5.645.440,00
revenue	Spain	114	24.637.670,68	48.854.337,49	4.575.627,93
revenue	Belgium	119	18.498.986,82	40.824.950,76	3.742.417,10
revenue	**India**	**570**	**17.636.054,61**	**39.107.746,66**	**1.638.043,25**
revenue	**Russia**	**185**	**8.022.980,58**	**14301548,6**	**1051470,765**

Another significant association confirmed by the ANOVA is the one between the **revenues** and the **main production country** of the movie. As a matter of fact, **Chinese** and **UK** movies have **higher mean values** of revenues with respect to the rest of the world. On the other hand, **Indian** and **Russian** movies are **not efficient** considering this **monetary feature**. The mean revenues of a **Chinese** movie are around **15 times higher** than a **Russian** one. The F-test is high, with a value of 22.966. (Levene-test=41.423)

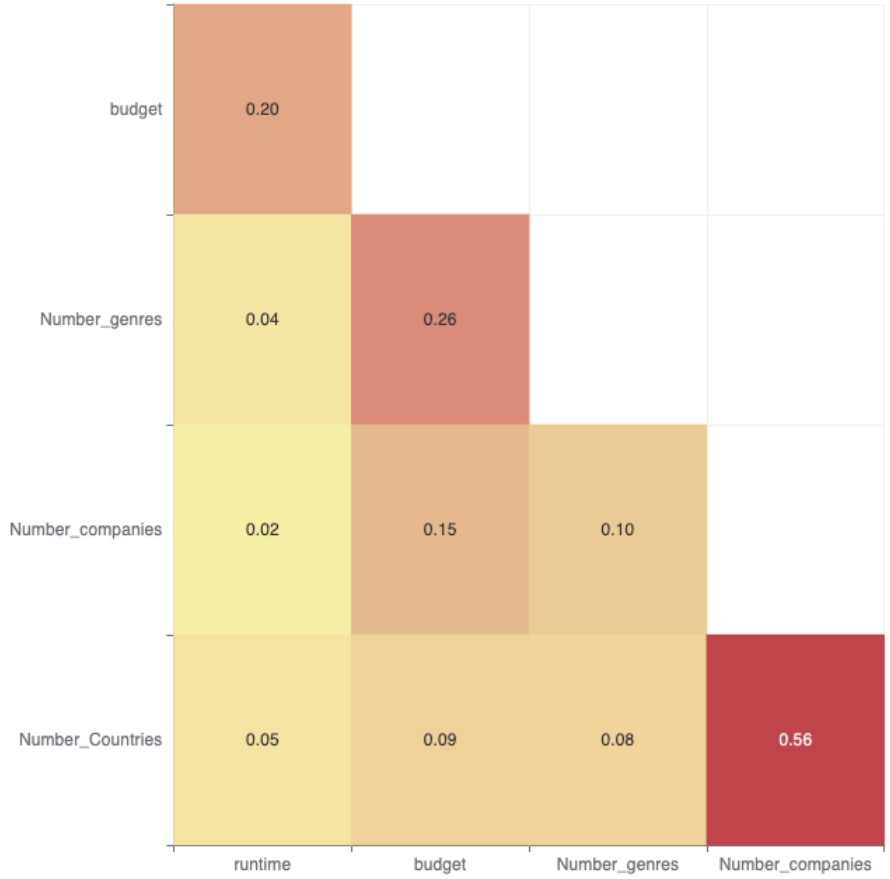
Revenue – Main Company Size

	Group	N	Mean	Std. Deviation	Std. Error
revenue	**big**	**1547**	**119000000**	**208000000**	**5277685,551**
revenue	medium	2008	110000000	222000000	4943015,142
revenue	**small**	**5872**	**43823948,2**	**105000000**	**1370884,38**

Again, the ANOVA shows that the **mean revenues** of different groups of **company size are significantly different.** In fact, as we expected, the mean revenues for the **biggest main producers** are **higher** than the lower ones. The F-test (225.387) shows the significancy of the analysis, and also the Levene-test value confirms this statement. As a result, we can affirm that the revenues of a big main company producing a movie are almost 3 times the incomes of a small company.

II. Data Preparation – *Bivariate Analysis*

Then we checked for redundancy among the quantitative explanatory variables, since multicollinearity can cause difficulties when interpreting the effects of the individual predictors. By performing bivariate analyses between explanatory variables, we can also find out some interesting patterns that can't be observed by analyzing the predictors individually.



As we can notice from the heatmap, there's no significantly high correlation between our quantitative explanatory variables, except for "*Number_Countries*" and "*Number_Companies*" which are positively correlated (0.56) with level of significance at 0.01.

Movies that are produced in a lot of different countries also require the participation of a lot of different companies.

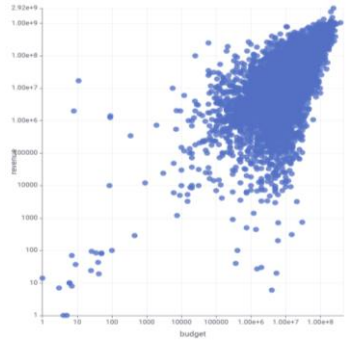
This will be taken into account when performing the logistic regression.

Before computing the correlations, we expected the budget to be highly correlated with the runtime. We thought that longer movies would be more expensive.
By looking at the coefficient value (0.2), we can see that they are positively correlated, but the intensity of the association is not so strong.

II. Data Preparation – *Bivariate Analysis*

Budget

Budget - Revenue



A correlation coefficient of **0.74** together with a p-value equal to 0 suggest that **higher budgets are associated with higher revenues**.
Since p-value = 0, we visually inspected the scatter plot distribution, which also illustrates the upward trend between the two variables.
(The scatter-plot was plotted by taking the log-transformed values of budget and revenue)

Budget - Genre

	Group	N	Mean	Std. Deviation	Std. Error
budget	**Adventure**	557	**53.374.255,42**	63.644.670,13	2.696.711,22
budget	**Animation**	280	**52.628.042,96**	54.961.026,45	3.284.549,56
budget	**Science Fiction**	223	**49.452.016,62**	63.472.727,97	4.250.448,25
budget	**Fantasy**	205	**47.153.106,26**	57.690.297,51	4.029.265,16
budget	**Family**	191	**44.794.532,40**	49.475.273,05	3.579.905,12
budget	Action	1451	39.567.780,09	52.567.134,69	1.380.004,81
budget	War	93	27.391.804,40	39.422.289,05	4.087.900,94
budget	Total	9427	24.993.007,82	38.702.791,52	398.616,74
budget	Mystery	120	24.068.637,90	43.015.016,58	3.926.715,82
budget	Thriller	417	20.906.286,58	26.440.817,36	1.294.812,40
budget	History	62	20.670.680,37	24.438.440,40	3.103.685,03
budget	Crime	440	18.360.857,48	23.719.886,41	1.130.801,22
budget	Comedy	2045	16.783.136,30	20.667.981,52	457.037,05
budget	Drama	2181	15.303.686,56	21.137.429,42	452.610,20
budget	Western	68	13.840.062,41	22.951.334,07	2.783.258,08
budget	Romance	276	13.707.093,11	17.274.442,91	1.039.799,27
budget	Music	96	12.577.432,61	16.413.926,31	1.675.239,34
budget	Horror	621	11.437.725,86	16.768.504,07	672.896,89
budget	**Documentary**	98	**2.331.924,76**	4.964.214,73	501.461,41
budget	**TV Movie**	3	**1.700.000,33**	2.858.320,89	1.650.252,33

The ANOVA table shows that, on average, **adventure**, **animation**, **sci-fi**, **fantasy** and **family** movies tends to have the highest budgets, while **documentaries** and **TV movies** tends to have the lowest budgets.
A significant F-test value of **80.03** indicates a substantial difference between group means compared to the variation within groups.
We also looked at the Levene-test since budget is not normally distributed (Levene-test = 143.5)

Budget – Production Country

	Group	N	Mean	Std. Deviation	Std. Error
budget	**China**	148	**47.015.210,18**	56.349.761,56	4.631.921,87
budget	**Germany**	277	**33.409.234,88**	35.104.633,67	2.109.233,02
budget	**UK**	583	**30.391.963,53**	46.998.290,19	1.946.470,30
budget	**USA**	5192	**28.841.878,17**	41.727.638,70	579.103,87
budget	**Canada**	400	**28.737.040,36**	37.840.399,32	1.892.019,97
budget	Australia	163	24.995.903,01	32.247.843,97	2.525.846,08
budget	Total	9427	24.993.007,82	38.702.791,52	398.616,74
budget	Japan	167	19.574.271,71	32.564.241,89	2.519.896,69
budget	France	463	19.026.657,41	23.170.549,37	1.076.826,75
budget	Belgium	119	16.093.863,43	22.218.884,11	2.036.801,77
budget	Spain	114	15.680.783,41	26.045.135,99	2.439.350,48
budget	other	933	15.185.499,75	31.287.183,20	1.024.296,39
budget	**Italy**	113	**13.836.513,22**	21.724.588,79	2.043.677,40
budget	**India**	570	**6.834.293,05**	14.465.650,15	605.899,41
budget	**Russia**	185	**5.154.658,75**	7.027.131,68	516.645,00

For movies where the main production country are **China**, **Germany**, **UK**, **USA** or **Canada** tends to have, on average, the highest budget. On the other hand, movies produced in **Russia**, **India** and **Italy** have the lowest budgets.
Countries with the highest budget have, on average, **4 times higher budget** than the countries with the lowest budgets.
(Levene-test = 42.7)

Budget – Company Size

	Group	N	Mean	Std. Deviation	Std. Error
budget	big	1547	38.094.102,10	50.093.547,60	1.273.610,50
budget	medium	2008	34.768.216,30	45.557.591,21	1.016.667,40
budget	small	5872	18.198.727,93	30.175.228,94	393.783,53

Finally, this correlation between budget and company size is quite straight-forward: **bigger companies have, on average, higher budgets**.

In this case, a significant Levene-test equal to 277.8 implies an even bigger difference between groups vs within groups.

Now, does a higher budget make a movie more successful? Are big budget movies «too big to fail»?

II. Data Preparation – *Bivariate Analysis*

Now that we saw which genres have, on average, the highest budgets, we wanted to find out some other interesting patterns regarding movie genres.

Genre

Genre - Runtime

	Group	N	Mean	Std. Deviation	Std. Error
runtime	**History**	62	**132**	27,7151	3,5198
runtime	**War**	93	**123**	24,2606	2,5157
runtime	**TV Movie**	3	**123**	29,9388	17,2852
runtime	Western	68	117	22,8125	2,7664
runtime	Drama	2181	117	23,144	0,4956
runtime	Romance	276	115	24,2311	1,4585
runtime	Action	1451	114	23,0902	0,6062
runtime	Adventure	557	113	23,2408	0,9847
runtime	Crime	440	113	20,3522	0,9703
runtime	Science Fiction	223	110	20,512	1,3736
runtime	Thriller	417	109	17,8977	0,8765
runtime	Fantasy	205	108	20,6873	1,4449
runtime	Mystery	120	108	17,7191	1,6175
runtime	Music	96	106	19,4399	1,9841
runtime	Comedy	2045	103	16,7134	0,3696
runtime	**Documentary**	98	**99**	28,8657	2,9159
runtime	**Horror**	621	**97**	13,8313	0,555
runtime	**Family**	191	**96**	15,0009	1,0854
runtime	**Animation**	280	**91**	12,0064	0,7175
runtime	Total	9427	109,5318	21,8986	0,2255

History, **war** and **TV movies** have, on average, the longest runtime.
Documentaries, **horror**, **family** and **animation** movies have the shortest runtime on average.
A significant F-test value of **80.03** indicates a substantial difference between group means compared to the variation within groups.

Genre – Number of Genres

	Group	N	Mean	Std. Deviation	Std. Error
Number_genres	**Animation**	280	**3,6286**	1,0931	0,0653
Number_genres	**Family**	191	**3,3089**	1,135	0,0821
Number_genres	**Adventure**	557	**3,2944**	0,9423	0,0399
Number_genres	**Fantasy**	205	**3,1707**	0,8829	0,0617
Number_genres	**War**	93	**3,0753**	1,0858	0,1126
Number_genres	Action	1451	2,9959	0,9344	0,0245
Number_genres	Mystery	120	2,9917	0,7724	0,0705
Number_genres	Science Fiction	223	2,9193	0,8609	0,0577
Number_genres	Crime	440	2,8136	0,7968	0,038
Number_genres	History	62	2,629	0,7517	0,0955
Number_genres	Thriller	417	2,5683	1,0028	0,0491
Number_genres	Romance	276	2,4746	0,7002	0,0421
Number_genres	Music	96	2,4479	1,045	0,1067
Number_genres	Horror	621	2,2287	0,8147	0,0327
Number_genres	Comedy	2045	2,1452	0,8995	0,0199
Number_genres	Drama	2181	2,138	0,9265	0,0198
Number_genres	**TV Movie**	3	**2**	1	0,5774
Number_genres	**Western**	68	**1,9118**	1,0613	0,1287
Number_genres	**Documentary**	98	**1,5408**	0,8754	0,0884
Number_genres	Total	9427	2,5339	1,031	0,0106

Crossover between genres can often give a boost to the quality of a movie and attract different types of audience.
With this correlation test, we wanted to see which genres are more synergistic with other genres.
Apart from animation movies, which is quite obvious by definition, **family**, **adventure**, **fantasy** and **war** movies are the most suitable for crossovers.

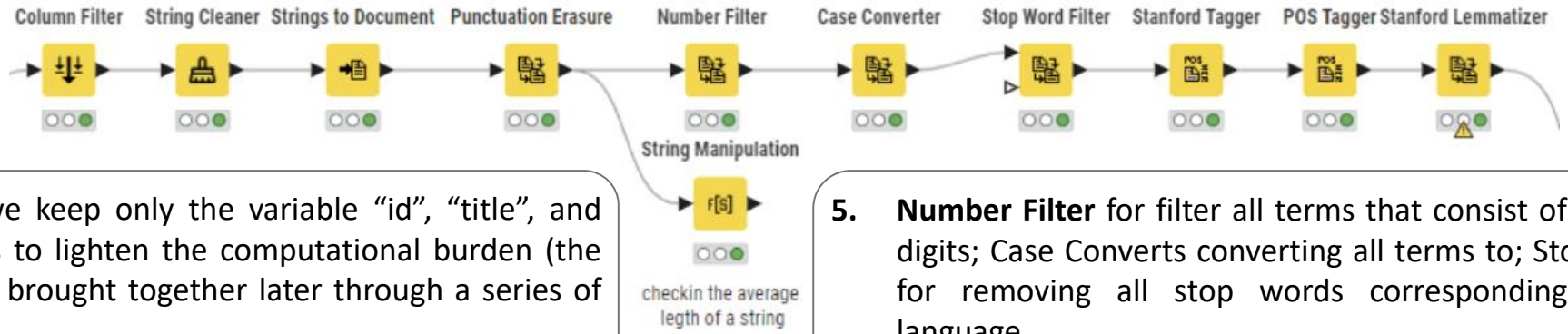
Genre – Number of Prod.Companies

	Group	N	Mean	Std. Deviation	Std. Error
Number_companies	**History**	62	**3,6452**	2,7999	0,3556
Number_companies	**Fantasy**	205	**3,6439**	2,7162	0,1897
Number_companies	**War**	93	**3,5806**	2,8297	0,2934
Number_companies	**Thriller**	417	**3,5324**	2,5924	0,127
Number_companies	**Animation**	280	**3,4393**	2,5109	0,1501
Number_companies	Adventure	557	3,3698	2,3259	0,0986
Number_companies	Crime	440	3,3091	2,5799	0,123
Number_companies	Science Fiction	223	3,296	2,1144	0,1416
Number_companies	Action	1451	3,295	2,3693	0,0622
Number_companies	Drama	2181	3,2682	2,5723	0,0551
Number_companies	Mystery	120	3,2417	2,2454	0,205
Number_companies	Horror	621	3,2383	1,948	0,0782
Number_companies	Family	191	3,2042	2,1825	0,1579
Number_companies	Documentary	98	2,7653	3,2389	0,3272
Number_companies	Comedy	2045	2,7467	1,9067	0,0422
Number_companies	Romance	276	2,6775	2,1065	0,1268
Number_companies	Music	96	2,6354	1,8129	0,185
Number_companies	Western	68	2,6324	1,9232	0,2332
Number_companies	TV Movie	3	2	0.0	0.0
Number_companies	Total	9427	3,1607	2,3392	0,0241

We wanted to see if some genre require multiple production companies to be involved.
There is statistical significance (F-test = **432.7**) that, on average, **history**, **fantasy**, **war**, **thriller** and **animation** movies have the highest number of companies involved in the production of the movie.
These genres probably require more domain knowledge than the others.

II. Data Preparation – *Text Normalization*

As mentioned earlier, within the dataset is the variable "**overview**" which provides a **brief description or summary** of the movie. We decided to take advantage of this variable and apply a text analysis to define the presence of potential Unigrams and Bigrams relevant in the definition of the film, in order thus to increase the accuracy and interpretability of the model.

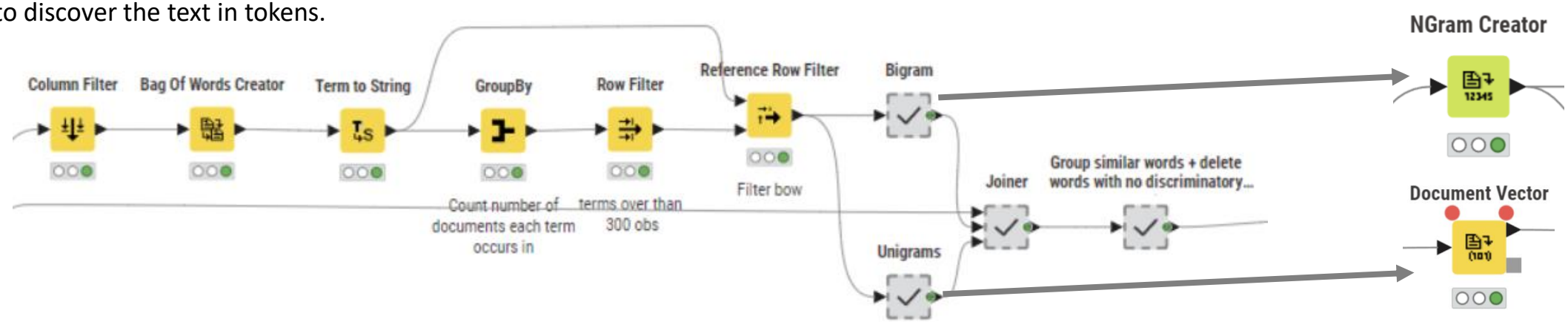


1. **Column Filter**: we keep only the variable “id”, “title”, and “overview” so as to lighten the computational burden (the workflow will be brought together later through a series of Joiner).
2. **String Cleaner**: we apply this node, to the “overview” that provides simple basic string cleaning operations like removing whitespace, removing punctuation or padding.
3. **String to Document**: we transform the “overview” into a document (essential for text cleaning and processing).
4. **Punctuation Erasure**: in this way we remove all punctuation characters of terms contained in the input documents.
5. **String Manipulation**: we calculate the length of each document. we want to verify that there are no documents that have a length that is not significant to be considered in the text analysis. the minimum length is 30 terms, which is sufficient to obtain useful insights (for this reason we do not apply "N Chart Filter“).

5. **Number Filter** for filter all terms that consist of numbers and digits; Case Converts converting all terms to; Stop Word Filter for removing all stop words corresponding the English language.
6. **Stanford Lemmatizer**: returns the lemma of a term by removing inflections, e.g in case of plurals, pronoun case, and verb endings. It results to be more precise than the Snowball stemmer (that cut very often incorrectly the root of many words) because the lemma is based heavily on the Part-Of-Speech (POS) tag of a term. The POS tagger is applied to each terms before, using the nodes **Stanford Tagger** and **POS Tagger**. These two nodes combined assign a label to each word to indicate the part of speech and often also other grammatical categories; in this way the Stanford Lemmatizer will return the word lemma based on the assigned tag.

II. Data Preparation – *Unigrams & Bigrams*

now that the overview has been cleaned through text normalization and represented in the new variable "Preprocessed_document," we can proceed to discover the text in tokens.



1. We apply the **Bags of Words** node to uncover each document in the terms from which it is composed
2. We consider only the most frequent terms, which **have more than 300 observations**.
3. Unigrams: with **Document Vector** node we create a dummy for each terms selected (1= present in the 0= absent in the overview).
4. Bigrams: with **Ngram Creator** node we create Bigrams associated with the most frequent terms. Then we slect only the Bigrams which **have more than 40 observations**. we create a dummy for each Bigrams selected (1= present in the 0= absent in the overview).
5. Through a **joiner** we join Unigrams and Bigrams.
6. Through manual work we examine the final Unigrams and Bigrams going to **eliminate the non-discriminatory** Ngrams to explain the content of a movie.
7. In addition, still manually, we grouped Unigrams and/or Bigrams with **similar meanings** within a single variable.
8. After this text processing we obtained **42 discriminatory variables** between Unigrams and Bigrams.

world_war
woman
true story
true love
town
time
tell story
team
secret
school
save
san francisco
road trip
return home
return
police officer
past

mysterious
meet
marry
love life
love
lo angele
live
life
killer_terms
join force
help
friend
fight
family_memb...
fall love

fall
escape
dream
discover
difficulties
death
city
child
american
New York


MODELS
EVALUATION

III. Models Evaluation

Once we have prepared the dataset and defined our target variable (successful_movie [0-1]), we can proceed to setup the train-test splitting process.

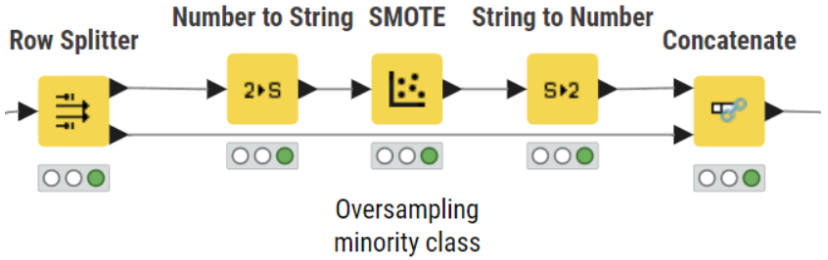
1. **Column Filter**: we started by filtering out all irrelevant variables in the definition of the target variable:
   - "budget" → replaced with "budget_dummy" variable
   - "vote_average", "vote_count", "weighted_rating"→ all these variables were components for the calculation of the new target variable "successful_movie"
   - "revenue" (same reason explained in the building target slide)
   - "release_date", "Year" , "id", "title" → they are not functional in defining a film's success within our analysis
   - "production_companies" "production_countries" → replaced with the dummy for the most relevant countries/companies and the dummy for non-relevant countries/companies
   - "Main Genre" → replaced as its usefulness was only for the purposes of Bivaraites
   - "overview", "Document" → replaced with dummies for Unigrams, Bigrams, and more relevant terms
2. **Partitioning**: we divided the film sample into training sets (70% of the observations) and test set (30% observations) utilizing the method "Drawn randomly".
3. **Class Imbalance**: through the Value counter node we checked for the presence of any Class imbalance for the target variable in the training test. from the output emerges an effective with a ratio of **80 to 20** against the minority classes.

before

0	5369
1	1229

after

0	5369
1	3687



4. **Oversampling minority class**: considering the class imbalance, we decided to oversample the minority class in order to increase the predictive effectiveness of the models. On the other hand, we avoided deviating too much from the actual ratio and thus maintaining a ratio of **60 to 40** against of the minority class**.** We used the following nodes.

- **Row Splitter**: we divided the training test based on the target variable, in this way only the observation belonging to the target class are oversampled.
- **SMOTE**: we applied the oversampling technique by a factor of 2 (we introduced two more portions for the target class) using the value 5 as the default choice for the "Nearest Neighbors" option.
  - **Number to String, String to Number** → we used these nodes to avoid that the Smote node would transform categorical variables (signed as integer) into float.
- **Concatenate**: we concatenated the oversampled target class with the unaltered non-target class obtaining our final training set.

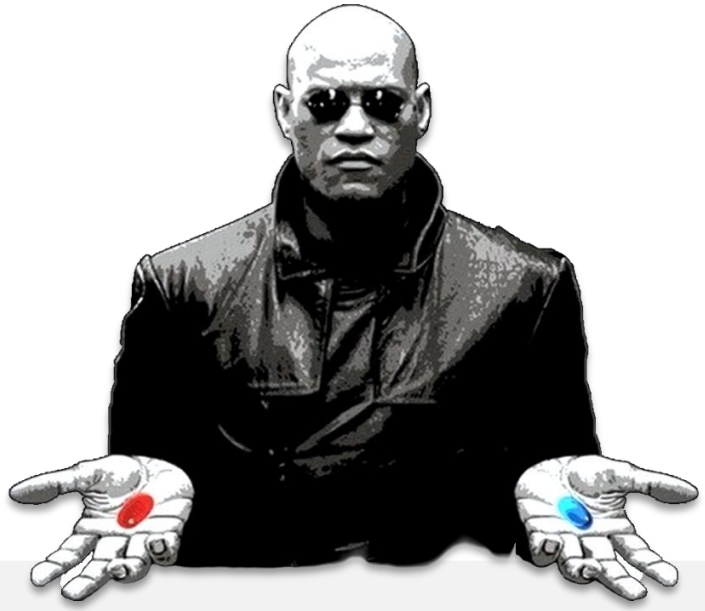
III. Models Evaluation – *Evaluation Criteria*

We first wanted to consider the trade-off between false positives and false negatives in terms of costs and risks.

False Positives

- Movies that are not successful but that are predicted to be successful -

This might lead to **misallocation of resources**, such as investing in marketing campaigns that ultimately fails to perform well, resulting in waste of time, money and effort.



False Negatives

- Movies that are successful but that are predicted not to be successful -

This might lead to **missed opportunities** like failing to buy licence for a movie that could have been profitable, resulting in revenue loss.

To determine which type of error is worse we would have to consider company's objectives and priorities:

- If the goal is to **minimize financial losses** and avoid investing in unsuccessful movies, then false positives can be considered worse
- If the goal is to **maximize revenues**, then false negatives can be considered worse

In addition, company's strengths and weaknesses should be considered, such as the **ability to mitigate losses** and **risk management strategies**.

Given the complexity of the identification of specific revenues and advertising spending, which are essential for quantifying potential revenues and losses, we aimed at finding an **optimal balance** between minimizing false positive and false negative.
Our classification models will be evaluated by considering the **Youden's Index** to automatically spot the optimal threshold, along side the ROC curve, and also by comparing the **AUC**, which is not affected by the threshold.

III. Models Evaluation – *Models Overview*

Parameter optimization loop set always to maximize Youden Index

In order to obtain the best prediction based on the criteria expressed above, we decided to run several classification models with different combinations. let's start by analysing the applied tree models.

Classification Trees

- Gain Ratio without pruning
- Gain Ration with MDL pruning method
- Gini Index without pruning
- Gini Index with MDL pruning method

We implemented 4 models based on the application or non-application of pruning and the splitting rule used (Gini index or Gain Ratio).
For each combination we applied Parameter Optimization for:

- minimum number of records per node
- maximum number of nominal values.

We used the Brute Force strategy.

Random Forest

- Information Gain splitting rule
- Information Gain Ratio splitting rule
- Gini Index splitting rule

We implemented 3 models based on the the splitting rule used (Information Gain, Information Gain Ratio, Gini Index).
For each combination we applied Parameter Optimization for:

- Number of models
- Minimum node size
- Maximum number of levels (tree depth)
- Best learning rate

We used Bayesian Optimization strategy in order to balance the trade off between computational resources and best parameters obtained.

XGBoost

- XGB with parameter optimization loop
- XGB with feature selection loop

We implemented 2 models one with just parameter optimization and and one with the addition of feature selection.
Parameter Optimization for:

- Maximum number of levels (tree depth)
- Minimum number of observations

We used Bayesian Optimization strategy for computational reason
The Feature Selection is applied using the Forward strategy and a threshold for number of feature equal to 25.

III. Models Evaluation – *Models Overview*

Parameter optimization loop set always to maximize Youden Index

We also applied the following models:

Logistic Regression

Newton-Raphson algorithm with uniform method

SGD algorithm with uniform method

SGD algorithm with Laplace (L1 regularization)

SGD algorithm with Gauss (L2 regularization)

We implemented 4 models based on the choice of algorithm (Newton-Raphson or Stochastic Average Gradient) and the eventual usage of Regularization.
Using the SGD algorithm we normalized both training and test set inside the range of 0-1

Parameter Choice

- Both for Laplace and Gauss we set a **Variance** equal to 0,1
- **Maximum number of epoch** of 10 k in order to balance the trade off between computational resources best model

MLP

Hyperparameter optimization loop

Before applying the model we normalized both training and test set using the Z-Score Normalization
We applied the Parameter Optimization for:

- Number of layers
- Number of units

We used the Brute Force strategy
Best combination suggested by the optimization loop 2 layers with 30 neurons.

Cart Algorithm

SimpleCART (3.7) + Weka

We also implemented this simplified version of the C4.5 algorithm for **easier interpretation** of the results, although it doesn't perform as good as decision trees or random forests.
Before applied the model we normalized both training and test set inside the range of 0-1.
We set the minimumnumber of observations in a leaf node equal to 30 and we adopted a pruning procedure

III. Models Evaluation – *Models Comparison*

We concatenated the results of each individual model into a single table.

Model Name	Youden	Sensitivity	Specificity	AUC	Accuracy	Precision
RF - Information Gain	**0,461**	0,704	0,757	**0,803**	0,748	0,381
RF - Gini Index	0,459	0,669	0,790	0,801	0,769	0,404
RF - Information Gain Ratio	0,458	0,730	0,728	0,797	0,728	0,363
LOGIT - SGD + Gauss (L2)	0,453	0,766	0,687	0,785	0,701	0,342
LOGIT - Least Square	0,451	0,760	0,691	0,784	0,703	0,343
GB + Param. Opt. Loop	0,445	0,774	0,671	0,803	0,689	0,334
LOGIT - SGD + Laplace (L1)	0,445	0,748	0,697	0,783	0,706	0,344
DT - Gini Index + MDL	0,440	0,690	0,751	0,775	0,740	0,370
LOGIT Uniform	0,429	0,710	0,720	0,780	0,718	0,350
DT - Gain Ratio + MDL	0,425	0,663	0,762	0,761	0,744	0,372
WEKA	0,425	0,718	0,707	0,756	0,709	0,342
GB + Feature	0,418	0,643	0,775	0,769	0,752	0,378
DT - Gini Index + No Pruning	0,394	0,571	0,824	0,750	0,779	0,408
DT - Gain Ratio + No Pruning	0,380	0,726	0,654	0,754	0,667	0,308
MLP	0,294	0,611	0,683	0,697	0,671	0,291

Since we're not able to precisely quantify potential **losses** of false positives and potential **missed revenues** of false negatives, we prioritized the maximization of **Youden's Index**.

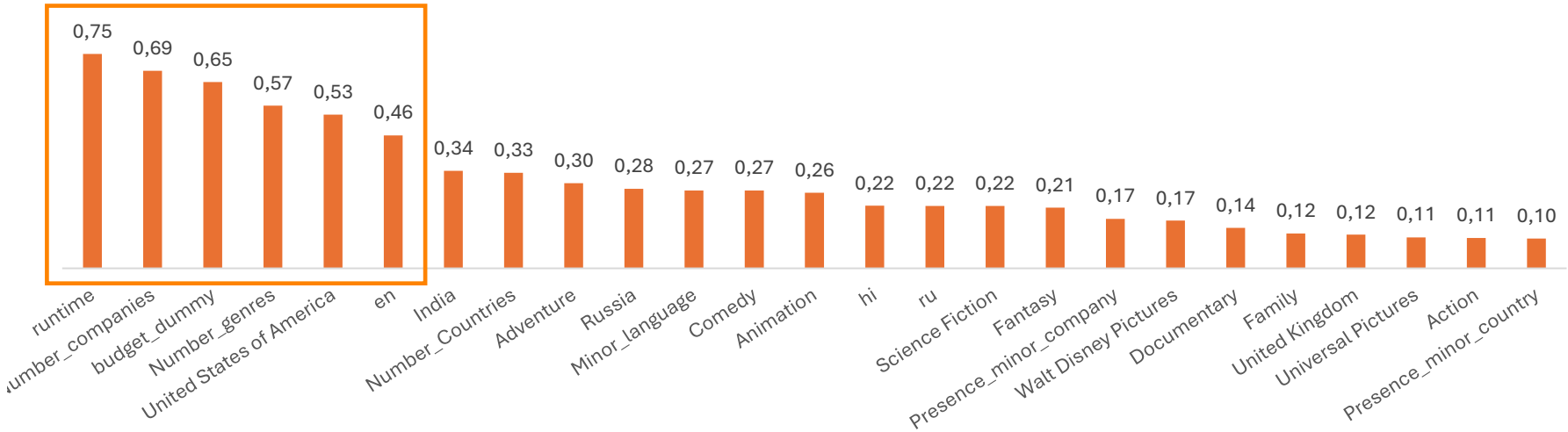
Random Forest with the **information gain** splitting rule, with a Youden's Index equal to **0.461**, is the model that finds the best balance between specificity and sensitivity.
Although it's far from 1, a Youden's Index of 0.461 is still significantly higher than 0, suggesting that the model has a **relatively high level of discriminative ability**.
Random Forest with information gain is the best model also according to the **AUC**, joined by the **Grandient Boosting** model with parameters optimization loop, both scoring **0.803**.
Unfortunately, GB scored worse in all the other metrics.

If we were to consider, instead, the Accuracy as our main performance indicator, the **Decision Tree** model with **Gini Index** splitting rule and **without pruning** would have been the best choice. Also, if we knew that potential losses derived from false positives are worse than missed revenues from false negatives, we would have also chosen the Decision Tree with Gini Index and without pruning as our best model, since it scored the highest **specificity** (**0.824**).

III. Models Evaluation – *Random Forest*

With the parameter optimization loop, we obtained the following combination of hyperparameters for the RF with Information Gain splitting rule:

Max Depth: **47**
Min Node Size: **16**
N. of models: **451**



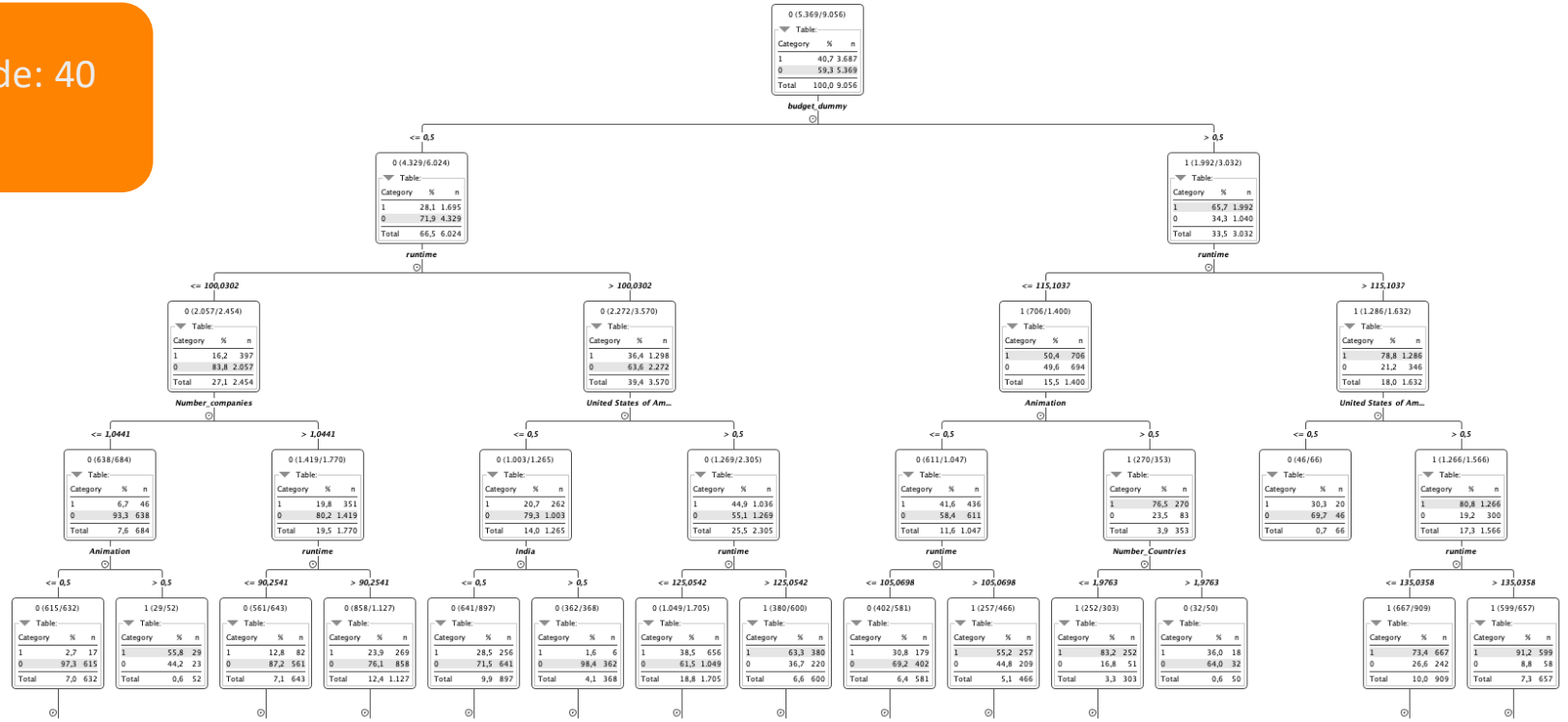
The **runtime**, the **number of companies** involved in the production of the movie, the **budget**, the **number of genres** a movie is classified as, the fact that the movie has been produced in the **US** and that the original language is **English** are the main variables that play a crucial role in determining whether a movie is going to be popular or not.
In particular, the runtime has the highest impact on a movie being successful. The importance value of 75% indicates that runtime is the significant predictor compared to the others.

Although Random Forest is the best performing model, since it's a "black-box" model, we decided to also explore other models, hoping to gain additional meaningful insights.

III. Models Evaluation – *Decision Tree (Gini, no pruning)*

Also in the case of the Decision Tree model, a parameter optimization loop has been applied:

minNumberRecordsPerNode: 40
maxNumNominalValues: 2



With the minimum number of records per node equal to 40, which is relatively high, we make the tree create larger nodes, leading to fewer splits in the tree. This should make the model less likely to be overfitting. Also, by setting the maximum number of nominal values to 2, we simplify the model by allowing two splits for each categorical variable.
We can compare the results from the Random Forest model with the decision tree. Here the first four levels of the tree are shown:
The budget has the highest predictive power and it suggests that movies with low budget are more likely to have low success.

III. Models Evaluation – *Logistic Regression*

Although **Logit models** performed worse than our best model (Random Forest with Information Gain), we still wanted to analyze the coefficients of the logistic regression models as they provide a more **straightforward interpretability** of the explanatory variables, both in terms of **direction** and **magnitude**. By doing this, we might discover further interesting insights that can improve the understanding of the previews models.

Model Name	Youden	Sensitivity	Specificity	AUC	Accuracy	Precision
RF - Information Gain	**0,461**	0,704	0,757	**0,803**	0,748	0,381
RF - Gini Index	0,459	0,669	0,790	0,801	0,769	0,404
RF - Information Gain Ratio	0,458	0,730	0,728	0,797	0,728	0,363
LOGIT - SGD + Gauss (L2)	0,453	0,766	0,687	0,785	0,701	0,342
LOGIT - Least Square	0,451	0,760	0,691	0,784	0,703	0,343
GB + Param. Opt. Loop	0,445	0,774	0,671	0,803	0,689	0,334
LOGIT - SGD + Laplace (L1)	0,445	0,748	0,697	0,783	0,706	0,344
DT - Gini Index + MDL	0,440	0,690	0,751	0,775	0,740	0,370
LOGIT Uniform	0,429	0,710	0,720	0,780	0,718	0,350
DT - Gain Ratio + MDL	0,425	0,663	0,762	0,761	0,744	0,372
WEKA	0,425	0,718	0,707	0,756	0,709	0,342
GB + Feature	0,418	0,643	0,775	0,769	0,752	0,378
DT - Gini Index + No Pruning	0,394	0,571	0,824	0,750	0,779	0,408
DT - Gain Ratio + No Pruning	0,380	0,726	0,654	0,754	0,667	0,308
MLP	0,294	0,611	0,683	0,697	0,671	0,291

We will focus on the logit model trained with **Stochastic Gradient Descent** algorithm and **Ridge Regularization**, reason being that it performed better, both in terms of **Youden's Index** and **AUC**, than the other three logit configurations.

When performing bivariate analysis, we found out a significant strong positive correlation between two explanatory variables, **Number of Companies** and **Number of Countries**, which might cause multicollinearity issues.

By applying a **Ridge Regularization**, the coefficients of correlated variables will be penalized and this will help reducing the impact of multicollinearity.

III. Models Evaluation – *Logit SGD L2 (Gauss)*

Variable	Coeff.	Std. Err.	z-score	P>|z|	odds_ratio	low_95%	upp_95%
runtime	4.274	0.193	22.13	0	71.791	3.895	4.652
Animation	1.316	0.125	10.565	0	3.728	1.072	1.56
budget_dummy	0.842	0.061	13.752	0	2.322	0.722	0.963
Constant	-3.789	0.204	-18.541	0	0.023	-4.189	-3.388
India	-1.339	0.173	-7.742	0	0.262	-1.678	-1
United States of America	0.818	0.106	7.719	0	2.265	0.61	1.025
United Kingdom	0.42	0.079	5.35	0	1.522	0.266	0.574
Presence_minor_company	0.564	0.106	5.321	0	1.757	0.356	0.771
Universal Pictures	0.475	0.093	5.099	0	1.608	0.292	0.657
Comedy	-0.365	0.073	-5.005	0	0.694	-0.508	-0.222
Walt Disney Pictures	0.71	0.146	4.864	0	2.034	0.424	0.996
life	0.313	0.066	4.738	0	1.367	0.183	0.442
school	0.501	0.108	4.639	0	1.65	0.289	0.712
Fantasy	0.405	0.091	4.461	0	1.499	0.227	0.583
Columbia Pictures	0.433	0.102	4.236	0	1.541	0.233	0.633
New Line Cinema	0.545	0.132	4.128	0	1.724	0.286	0.803
Metro-Goldwyn-Mayer	-0.532	0.14	-3.806	0	0.587	-0.807	-0.258
Documentary	-0.876	0.232	-3.783	0	0.416	-1.33	-0.422
Science Fiction	0.336	0.089	3.766	0	1.399	0.161	0.511
Paramount	0.35	0.102	3.417	0.001	1.419	0.149	0.55
zh	-0.784	0.252	-3.111	0.002	0.456	-1.278	-0.29
Miramax	0.457	0.154	2.968	0.003	1.579	0.155	0.759
Music	-0.409	0.139	-2.936	0.003	0.665	-0.682	-0.136
France	0.326	0.112	2.905	0.004	1.386	0.106	0.547
Number_companies	0.68	0.235	2.899	0.004	1.974	0.22	1.14
Warner Bros. Pictures	0.256	0.09	2.836	0.005	1.292	0.079	0.433
Lionsgate	0.437	0.155	2.812	0.005	1.548	0.132	0.741
Family	-0.304	0.112	-2.709	0.007	0.738	-0.523	-0.084
Russia	-0.685	0.254	-2.693	0.007	0.504	-1.183	-0.186
fight	0.29	0.109	2.669	0.008	1.336	0.077	0.503
true story	0.459	0.182	2.529	0.011	1.583	0.103	0.816
StudioCanal	-0.456	0.183	-2.489	0.013	0.634	-0.814	-0.097
Summit Entertainment	0.402	0.162	2.473	0.013	1.494	0.083	0.72
love	-0.252	0.102	-2.467	0.014	0.777	-0.452	-0.052
woman	-0.232	0.095	-2.437	0.015	0.793	-0.419	-0.045
Action	-0.187	0.078	-2.399	0.016	0.829	-0.34	-0.034
DreamWorks Pictures	0.387	0.165	2.342	0.019	1.472	0.063	0.71
Drama	0.168	0.072	2.324	0.02	1.183	0.026	0.31
dream	-0.321	0.139	-2.308	0.021	0.725	-0.594	-0.048
Canal+	0.383	0.169	2.272	0.023	1.467	0.053	0.714
team	-0.251	0.113	-2.225	0.026	0.778	-0.473	-0.03
Crime	0.177	0.082	2.161	0.031	1.193	0.016	0.337
return	-0.262	0.122	-2.145	0.032	0.77	-0.501	-0.023
san francisco	0.474	0.221	2.145	0.032	1.607	0.041	0.908
ja	0.468	0.223	2.095	0.036	1.596	0.03	0.905
20th Century Fox	0.209	0.101	2.065	0.039	1.233	0.011	0.408
Australia	-0.336	0.166	-2.023	0.043	0.715	-0.661	-0.01
South Korea	0.38	0.195	1.953	0.051	1.463	-0.001	0.762
en	0.302	0.156	1.942	0.052	1.353	-0.003	0.608
it	0.465	0.24	1.94	0.052	1.592	-0.005	0.935
Adventure	0.157	0.082	1.909	0.056	1.17	-0.004	0.319
United Artists	0.306	0.161	1.904	0.057	1.358	-0.009	0.621
Focus Features	0.329	0.18	1.827	0.068	1.39	-0.024	0.682
save	-0.206	0.113	-1.821	0.069	0.814	-0.427	0.016

Considering that a lot of variables have p-value = 0, it makes sense in this case to also consider the z-score. As we can see, all z-score associated with variables with p-value = 0 are, in absolute terms, greater than 2, meaning they are statistically significant.

Runtime

For every one-unit increase in the runtime, the odds of a movie being successful (compared to not being successful) are **71.8** times higher, holding all other variables constant.
This means that **longer runtimes** are associated with **higher odds of a movie being successful** compared to not being successful.
The high value of the odds-ratio, equal to 71.8, indicates that runtime is a **highly significant predictor** of the likelihood of a movie being successful.

Budget

The odds of a movie being successful are **2.3** times higher when it has high budget compared to when it has low budget, ceteris paribus.

III. Models Evaluation – *Logit SGD L2 (Gauss)*

Variable	Coeff.	Std. Err.	z-score	P>|z|	odds_ratio	low_95%	upp_95%
runtime	4.274	0.193	22.13	0	71.791	3.895	4.652
Animation	1.316	0.125	10.565	0	3.728	1.072	1.56
budget_dummy	0.842	0.061	13.752	0	2.322	0.722	0.963
Constant	-3.789	0.204	-18.541	0	0.023	-4.189	-3.388
India	-1.339	0.173	-7.742	0	0.262	-1.678	-1
United States of America	0.818	0.106	7.719	0	2.265	0.61	1.025
United Kingdom	0.42	0.079	5.35	0	1.522	0.266	0.574
Presence_minor_company	0.564	0.106	5.321	0	1.757	0.356	0.771
Universal Pictures	0.475	0.093	5.099	0	1.608	0.292	0.657
Comedy	-0.365	0.073	-5.005	0	0.694	-0.508	-0.222
Walt Disney Pictures	0.71	0.146	4.864	0	2.034	0.424	0.996
life	0.313	0.066	4.738	0	1.367	0.183	0.442
school	0.501	0.108	4.639	0	1.65	0.289	0.712
Fantasy	0.405	0.091	4.461	0	1.499	0.227	0.583
Columbia Pictures	0.433	0.102	4.236	0	1.541	0.233	0.633
New Line Cinema	0.545	0.132	4.128	0	1.724	0.286	0.803
Metro-Goldwyn-Mayer	-0.532	0.14	-3.806	0	0.587	-0.807	-0.258
Documentary	-0.876	0.232	-3.783	0	0.416	-1.33	-0.422
Science Fiction	0.336	0.089	3.766	0	1.399	0.161	0.511
Paramount	0.35	0.102	3.417	0.001	1.419	0.149	0.55
zh	-0.784	0.252	-3.111	0.002	0.456	-1.278	-0.29
Miramax	0.457	0.154	2.968	0.003	1.579	0.155	0.759
Music	-0.409	0.139	-2.936	0.003	0.665	-0.682	-0.136
France	0.326	0.112	2.905	0.004	1.386	0.106	0.547
Number_companies	0.68	0.235	2.899	0.004	1.974	0.22	1.14
Warner Bros. Pictures	0.256	0.09	2.836	0.005	1.292	0.079	0.433
Lionsgate	0.437	0.155	2.812	0.005	1.548	0.132	0.741
Family	-0.304	0.112	-2.709	0.007	0.738	-0.523	-0.084
Russia	-0.685	0.254	-2.693	0.007	0.504	-1.183	-0.186
fight	0.29	0.109	2.669	0.008	1.336	0.077	0.503
true story	0.459	0.182	2.529	0.011	1.583	0.103	0.816
StudioCanal	-0.456	0.183	-2.489	0.013	0.634	-0.814	-0.097
Summit Entertainment	0.402	0.162	2.473	0.013	1.494	0.083	0.72
love	-0.252	0.102	-2.467	0.014	0.777	-0.452	-0.052
woman	-0.232	0.095	-2.437	0.015	0.793	-0.419	-0.045
Action	-0.187	0.078	-2.399	0.016	0.829	-0.34	-0.034
DreamWorks Pictures	0.387	0.165	2.342	0.019	1.472	0.063	0.71
Drama	0.168	0.072	2.324	0.02	1.183	0.026	0.31
dream	-0.321	0.139	-2.308	0.021	0.725	-0.594	-0.048
Canal+	0.383	0.169	2.272	0.023	1.467	0.053	0.714
team	-0.251	0.113	-2.225	0.026	0.778	-0.473	-0.03
Crime	0.177	0.082	2.161	0.031	1.193	0.016	0.337
return	-0.262	0.122	-2.145	0.032	0.77	-0.501	-0.023
san francisco	0.474	0.221	2.145	0.032	1.607	0.041	0.908
ja	0.468	0.223	2.095	0.036	1.596	0.03	0.905
20th Century Fox	0.209	0.101	2.065	0.039	1.233	0.011	0.408
Australia	-0.336	0.166	-2.023	0.043	0.715	-0.661	-0.01
South Korea	0.38	0.195	1.953	0.051	1.463	-0.001	0.762
en	0.302	0.156	1.942	0.052	1.353	-0.003	0.608
it	0.465	0.24	1.94	0.052	1.592	-0.005	0.935
Adventure	0.157	0.082	1.909	0.056	1.17	-0.004	0.319
United Artists	0.306	0.161	1.904	0.057	1.358	-0.009	0.621
Focus Features	0.329	0.18	1.827	0.068	1.39	-0.024	0.682
save	-0.206	0.113	-1.821	0.069	0.814	-0.427	0.016

Genres

Let's consider these movie genres: **Animation**, **Fantasy**, **Sci-fi**, **Drama** and **Crime**.
The odds of a movie being successful are higher when the movie belong to one of these genres compared to when it doesn't belong to them, ceteris paribus.
In particular, among all the genres, **animation** movies have the highest odds-ratio, equal to **3.7**.
Instead, for **Comedy**, **Documentary**, **Music**, **Family** and **Action** movies, the odds of a movie being successful are lower when the movie belong to one of these genres, ceteris paribus. The odds-ratio of these genres are all lower than 1.

Countries

When a movie is produced in **USA**, **UK** or **France**, the odds of them being successful are higher compared to when they are not produced in those countries, ceteris paribus, but the intensity of the association is not so strong (odds-ratio between **1.4** and **2.2**).
On the other hand, when a movie is produced in **India**, **Russia** or **Australia**, the odds of being successful is lower compared to when they are not produced in those countries, ceteris paribus.
It's also interesting to notice that **Japan**, **China**, **Canada**, **Spain**, **Italy**, **Germany** are not significant at all.

III. Models Evaluation – *Logit SGD L2 (Gauss)*

Variable	Coeff.	Std. Err.	z-score	P>|z|	odds_ratio	low_95%	upp_95%
runtime	4.274	0.193	22.13	0	71.791	3.895	4.652
Animation	1.316	0.125	10.565	0	3.728	1.072	1.56
budget_dummy	0.842	0.061	13.752	0	2.322	0.722	0.963
Constant	-3.789	0.204	-18.541	0	0.023	-4.189	-3.388
India	-1.339	0.173	-7.742	0	0.262	-1.678	-1
United States of America	0.818	0.106	7.719	0	2.265	0.61	1.025
United Kingdom	0.42	0.079	5.35	0	1.522	0.266	0.574
Presence_minor_company	0.564	0.106	5.321	0	1.757	0.356	0.771
Universal Pictures	0.475	0.093	5.099	0	1.608	0.292	0.657
Comedy	-0.365	0.073	-5.005	0	0.694	-0.508	-0.222
Walt Disney Pictures	0.71	0.146	4.864	0	2.034	0.424	0.996
life	0.313	0.066	4.738	0	1.367	0.183	0.442
school	0.501	0.108	4.639	0	1.65	0.289	0.712
Fantasy	0.405	0.091	4.461	0	1.499	0.227	0.583
Columbia Pictures	0.433	0.102	4.236	0	1.541	0.233	0.633
New Line Cinema	0.545	0.132	4.128	0	1.724	0.286	0.803
Metro-Goldwyn-Mayer	-0.532	0.14	-3.806	0	0.587	-0.807	-0.258
Documentary	-0.876	0.232	-3.783	0	0.416	-1.33	-0.422
Science Fiction	0.336	0.089	3.766	0	1.399	0.161	0.511
Paramount	0.35	0.102	3.417	0.001	1.419	0.149	0.55
zh	-0.784	0.252	-3.111	0.002	0.456	-1.278	-0.29
Miramax	0.457	0.154	2.968	0.003	1.579	0.155	0.759
Music	-0.409	0.139	-2.936	0.003	0.665	-0.682	-0.136
France	0.326	0.112	2.905	0.004	1.386	0.106	0.547
Number_companies	0.68	0.235	2.899	0.004	1.974	0.22	1.14
Warner Bros. Pictures	0.256	0.09	2.836	0.005	1.292	0.079	0.433
Lionsgate	0.437	0.155	2.812	0.005	1.548	0.132	0.741
Family	-0.304	0.112	-2.709	0.007	0.738	-0.523	-0.084
Russia	-0.685	0.254	-2.693	0.007	0.504	-1.183	-0.186
fight	0.29	0.109	2.669	0.008	1.336	0.077	0.503
true story	0.459	0.182	2.529	0.011	1.583	0.103	0.816
StudioCanal	-0.456	0.183	-2.489	0.013	0.634	-0.814	-0.097
Summit Entertainment	0.402	0.162	2.473	0.013	1.494	0.083	0.72
love	-0.252	0.102	-2.467	0.014	0.777	-0.452	-0.052
woman	-0.232	0.095	-2.437	0.015	0.793	-0.419	-0.045
Action	-0.187	0.078	-2.399	0.016	0.829	-0.34	-0.034
DreamWorks Pictures	0.387	0.165	2.342	0.019	1.472	0.063	0.71
Drama	0.168	0.072	2.324	0.02	1.183	0.026	0.31
dream	-0.321	0.139	-2.308	0.021	0.725	-0.594	-0.048
Canal+	0.383	0.169	2.272	0.023	1.467	0.053	0.714
team	-0.251	0.113	-2.225	0.026	0.778	-0.473	-0.03
Crime	0.177	0.082	2.161	0.031	1.193	0.016	0.337
return	-0.262	0.122	-2.145	0.032	0.77	-0.501	-0.023
san francisco	0.474	0.221	2.145	0.032	1.607	0.041	0.908
ia	0.468	0.223	2.095	0.036	1.596	0.03	0.905
20th Century Fox	0.209	0.101	2.065	0.039	1.233	0.011	0.408
Australia	-0.336	0.166	-2.023	0.043	0.715	-0.661	-0.01
South Korea	0.38	0.195	1.953	0.051	1.463	-0.001	0.762
en	0.302	0.156	1.942	0.052	1.353	-0.003	0.608
it	0.465	0.24	1.94	0.052	1.592	-0.005	0.935
Adventure	0.157	0.082	1.909	0.056	1.17	-0.004	0.319
United Artists	0.306	0.161	1.904	0.057	1.358	-0.009	0.621
Focus Features	0.329	0.18	1.827	0.068	1.39	-0.024	0.682
save	-0.206	0.113	-1.821	0.069	0.814	-0.427	0.016

Production Companies

Most of the famous production companies have positive, but not so strong values of odds-ratio, ranging between **1.2** and **1.7**. Only for **Walt Disney** the odds-ratio goes up to **2.03**, which is not intense, but still higher compared to the other production companies.

So the odds of a movie being successful is **2** times higher when it's produced by **Walt Disney** compared to when it's not produced by them, ceteris paribus.

On the other hand, the odds of a movie being successful is lower when it's produced by **Metro_Goldwyn_Mayer**, **Studio Canal** or other minor companies.

III. Models Evaluation – *Logit SGD L2 (Gauss)*

Variable	Coeff.	Std. Err.	z-score	P>|z|	odds_ratio	low_95%	upp_95%
runtime	4.274	0.193	22.13	0	71.791	3.895	4.652
Animation	1.316	0.125	10.565	0	3.728	1.072	1.56
budget_dummy	0.842	0.061	13.752	0	2.322	0.722	0.963
Constant	-3.789	0.204	-18.541	0	0.023	-4.189	-3.388
India	-1.339	0.173	-7.742	0	0.262	-1.678	-1
United States of America	0.818	0.106	7.719	0	2.265	0.61	1.025
United Kingdom	0.42	0.079	5.35	0	1.522	0.266	0.574
Presence_minor_company	0.564	0.106	5.321	0	1.757	0.356	0.771
Universal Pictures	0.475	0.093	5.099	0	1.608	0.292	0.657
Comedy	-0.365	0.073	-5.005	0	0.694	-0.508	-0.222
Walt Disney Pictures	0.71	0.146	4.864	0	2.034	0.424	0.996
life	0.313	0.066	4.738	0	1.367	0.183	0.442
school	0.501	0.108	4.639	0	1.65	0.289	0.712
Fantasy	0.405	0.091	4.461	0	1.499	0.227	0.583
Columbia Pictures	0.433	0.102	4.236	0	1.541	0.233	0.633
New Line Cinema	0.545	0.132	4.128	0	1.724	0.286	0.803
Metro-Goldwyn-Mayer	-0.532	0.14	-3.806	0	0.587	-0.807	-0.258
Documentary	-0.876	0.232	-3.783	0	0.416	-1.33	-0.422
Science Fiction	0.336	0.089	3.766	0	1.399	0.161	0.511
Paramount	0.35	0.102	3.417	0.001	1.419	0.149	0.55
zh	-0.784	0.252	-3.111	0.002	0.456	-1.278	-0.29
Miramax	0.457	0.154	2.968	0.003	1.579	0.155	0.759
Music	-0.409	0.139	-2.936	0.003	0.665	-0.682	-0.136
France	0.326	0.112	2.905	0.004	1.386	0.106	0.547
Number_companies	0.68	0.235	2.899	0.004	1.974	0.22	1.14
Warner Bros. Pictures	0.256	0.09	2.836	0.005	1.292	0.079	0.433
Lionsgate	0.437	0.155	2.812	0.005	1.548	0.132	0.741
Family	-0.304	0.112	-2.709	0.007	0.738	-0.523	-0.084
Russia	-0.685	0.254	-2.693	0.007	0.504	-1.183	-0.186
fight	0.29	0.109	2.669	0.008	1.336	0.077	0.503
true story	0.459	0.182	2.529	0.011	1.583	0.103	0.816
StudioCanal	-0.456	0.183	-2.489	0.013	0.634	-0.814	-0.097
Summit Entertainment	0.402	0.162	2.473	0.013	1.494	0.083	0.72
love	-0.252	0.102	-2.467	0.014	0.777	-0.452	-0.052
woman	-0.232	0.095	-2.437	0.015	0.793	-0.419	-0.045
Action	-0.187	0.078	-2.399	0.016	0.829	-0.34	-0.034
DreamWorks Pictures	0.387	0.165	2.342	0.019	1.472	0.063	0.71
Drama	0.168	0.072	2.324	0.02	1.183	0.026	0.31
dream	-0.321	0.139	-2.308	0.021	0.725	-0.594	-0.048
Canal+	0.383	0.169	2.272	0.023	1.467	0.053	0.714
team	-0.251	0.113	-2.225	0.026	0.778	-0.473	-0.03
Crime	0.177	0.082	2.161	0.031	1.193	0.016	0.337
return	-0.262	0.122	-2.145	0.032	0.77	-0.501	-0.023
san francisco	0.474	0.221	2.145	0.032	1.607	0.041	0.908
ja	0.468	0.223	2.095	0.036	1.596	0.03	0.905
20th Century Fox	0.209	0.101	2.065	0.039	1.233	0.011	0.408
Australia	-0.336	0.166	-2.023	0.043	0.715	-0.661	-0.01
South Korea	0.38	0.195	1.953	0.051	1.463	-0.001	0.762
en	0.302	0.156	1.942	0.052	1.353	-0.003	0.608
it	0.465	0.24	1.94	0.052	1.592	-0.005	0.935
Adventure	0.157	0.082	1.909	0.056	1.17	-0.004	0.319
United Artists	0.306	0.161	1.904	0.057	1.358	-0.009	0.621
Focus Features	0.329	0.18	1.827	0.068	1.39	-0.024	0.682
save	-0.206	0.113	-1.821	0.069	0.814	-0.427	0.016

Most Common Keywords in the Overview

Now we will analyze some of the most common keywords that we isolated from the overview of each movie by creating **unigrams** and **bigrams**.

It's interesting to notice that keywords having odds-ratio > 1 (*life, school, fight, true story, San Francisco*) could be considered as "***concrete***", while keywords having odds-ratio < 1 can be considered as "***abstract***" (*love, dream, team, return, woman* – in movies and, in general, in literature, women are often seen as inspiring muse).

So we can say that the odds of a movie being successful are higher, although not so impactful (odds-ratio between **1.3** and **1.7**), when the overview contains **concrete** words, compared to when a movie do not contain concrete words.

Instead, the odds of a movie being successful is lower when the overview contains abstract words compared to when a movie does not contain **abstract** Words.


TIME
SERIES
ANALYSIS

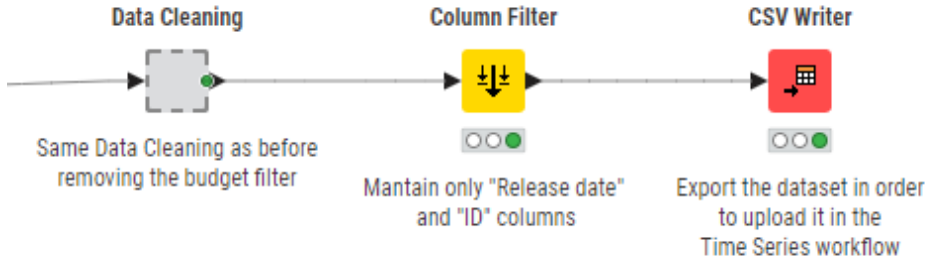
IV. Time Series Analysis – *Introduction*

The original uncleaned dataset is the same of the previous analysis. It represents a sample of 1,012,887 movies produced from the late 19th century to the present day in 2024, with daily release dates.

We proceeded in the study of our research question involving in the business case of our cinema a time series analysis. The aim of our cinema is to **estimate the number of films that will be produced each month** in order to **plan the logistics of its own halls and the right length of projection for each movie at the theater.** Looking at the management perspective of this research, we decided to analyse the time series that was produced by counting the **number of films that were released with monthly granularity starting from 2000.** In fact, this data represents both a traditionally used period of time and a turning point in the history of cinema, with the ultimate spread of the **digital cinema**. To make our dataset more reliable, we decided to use the data from the **previous analysis' data cleaning method**, and to export them through a brief series of nodes, culminating with the download of the CSV file. Briefly, we have decided to build a time series based on the dataset we had created for the previous analysis, counting the number of films released from a monthly perspective.

We have decided to examine a monthly granularity for two main reasons:

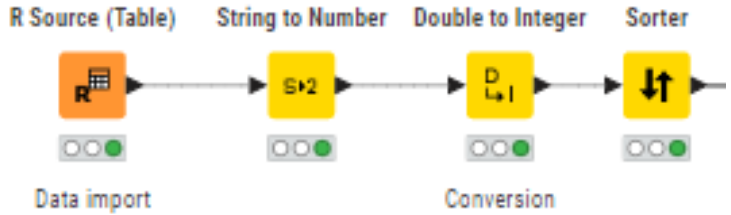
1. We had access to a **very large dataset** and were able to create a sufficiently long time series.
2. A **month** provides a **good approximation** of the time frame in which a movie is broadcasted in a theater hall.



On top, the image shows the process of data cleaning and manipulation which led us to the time series creation. The process was divided in three steps:

1. **Data cleaning metanode:** through this metanode, we used the same data cleaning process of the previous analysis, but we did not apply the budget filter because it was not significant to remove its missing values for our analysis. On the other hand, we decided to remove all the movies with revenues lower or equal to zero because they are not useful from our managerial point of view.
2. **Column filter:** this node allowed us to filter all the useless variables in order to keep only the "ID" and the "release_date" columns.
3. **CSV Writer:** we downloaded the "Dataset_TimeSeries" csv file before proceeding with the splitting of the dates, in order to do it with the R scripts in the second Knime workflow

IV. Time Series Analysis – *Data Preparation Overview*



The **"Movie_TimeSeries"** Knime workflow begins with the uploading of the **"Dataset_TimeSeries"** csv file. The R Source node is used in order to connect the R scripts and tools to the time series analysis. In this first application, we used the node to create the time series itself. In fact, the following Rstudio script was made in order to:

1. Separate the different components of the date format creating the columns **"Year"** and **"Month"** for each movie imported.
2. Count the **number of movies** which were released during each month of the **2000-2022 period.** The **2022** time series **limit** was imposed one main reason. As a matter of fact, the **118 days Hollywood strike** heavily affected the number of films produced in **2023** and **2024** in the **US**, where a high number of movies in out dataset were produced. In fact, we decided to **avoid new exogenous shocks** for our time series, since the **Covid-19** pandemic already had **notable effects** on our data.
3. Produce a dataset with the three columns useful to procede with our analysis, that are "Year", "Month" and "Movies". This last one columns, as a matter of fact, presents the count of the movies released with **monthly granularity.**

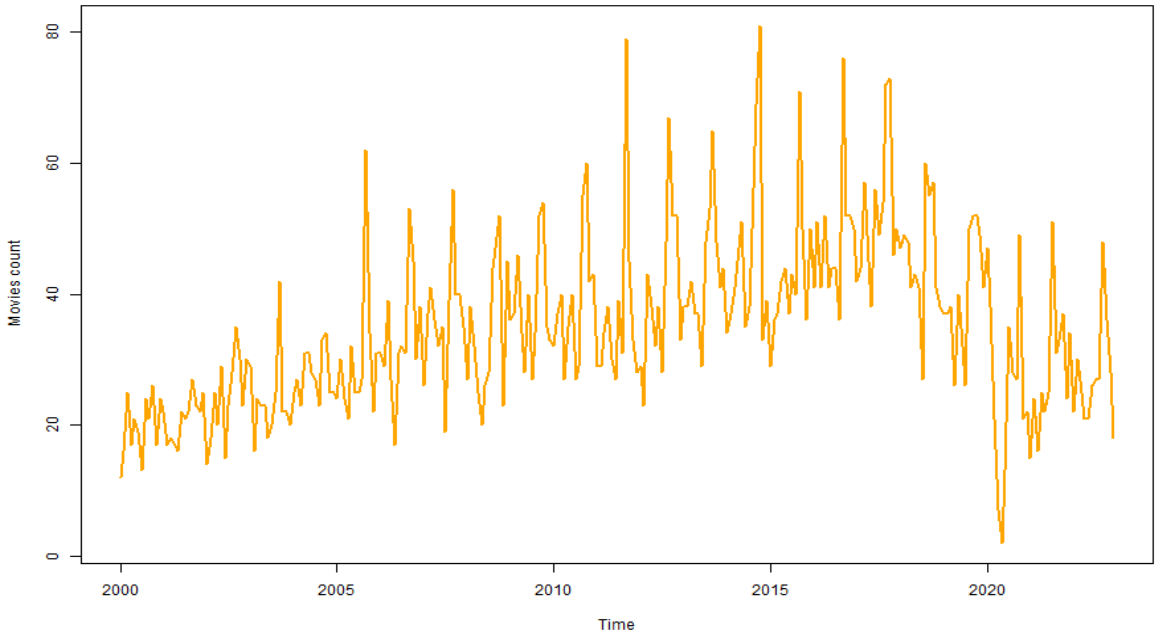
The "String to number" and the "Double to integer" nodes were useful to convert all the data of our three columns into integers. The "Sorter" node was used in order to **sort** our data in **two degrees**: firstly by **year** and secondly by **month.**

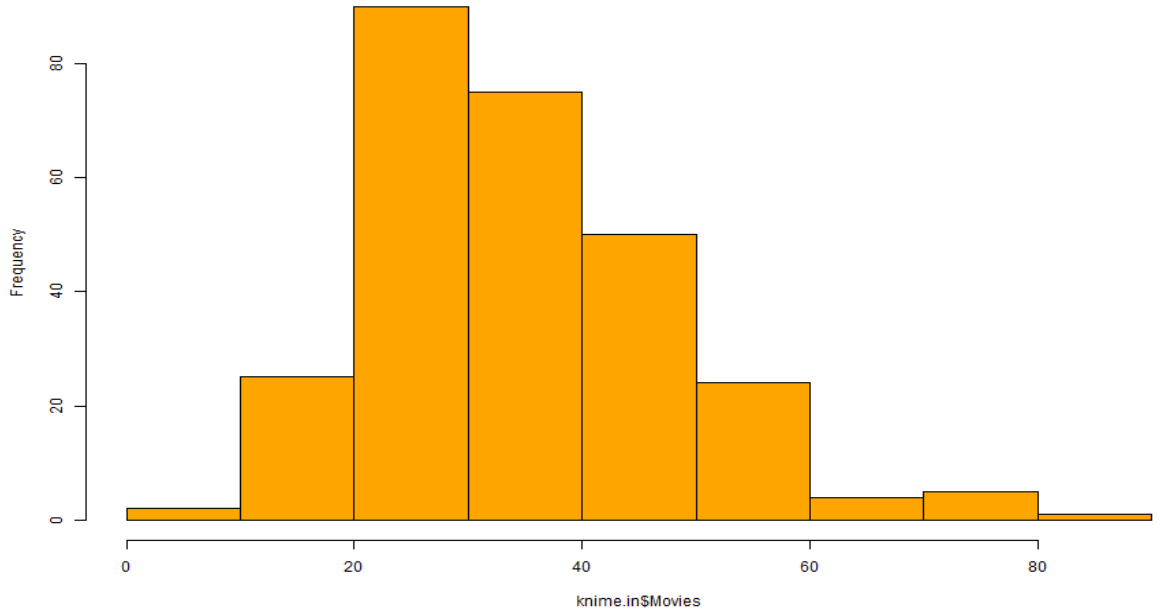
As a result, as we can see on the right, we obtained our cleaned and sorted time series, starting from **January 2000** until **December 2022** with **monthly granularity.**

Rows: 276 | Columns: 3

#	RowID	Year Number (integer)	Month Number (integer)	Movies Number (integer)
1	1	2000	1	12
2	24	2000	2	16
3	47	2000	3	25
4	70	2000	4	17
5	93	2000	5	21
6	116	2000	6	19
7	139	2000	7	13
8	162	2000	8	24
9	185	2000	9	21
10	208	2000	10	26
11	231	2000	11	17
12	254	2000	12	24
13	2	2001	1	22
14	25	2001	2	17
15	48	2001	3	18
16	71	2001	4	17
17	94	2001	5	16
18	117	2001	6	22
19	140	2001	7	21

IV. Time Series Analysis – *Exploratory Graphical Analysis (1)*

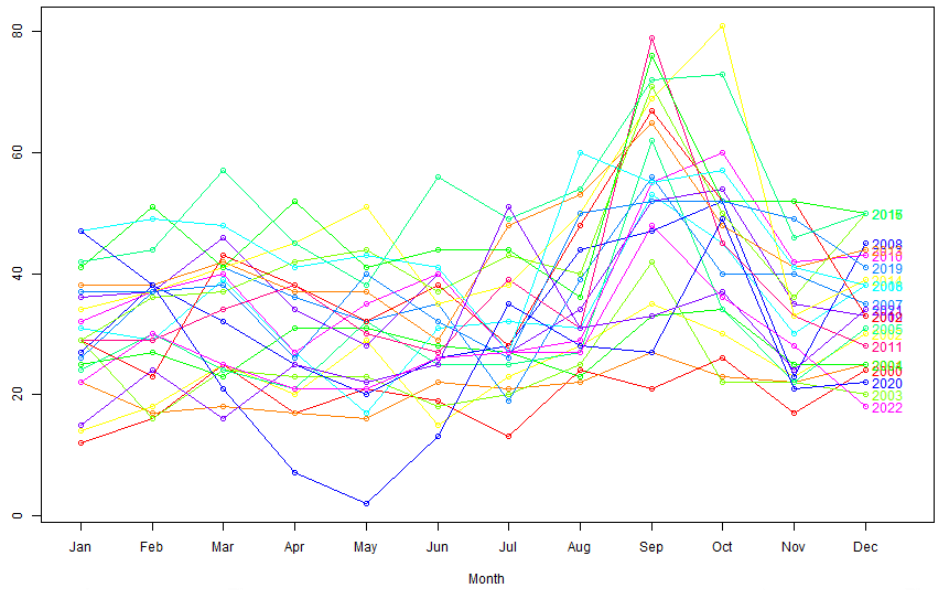


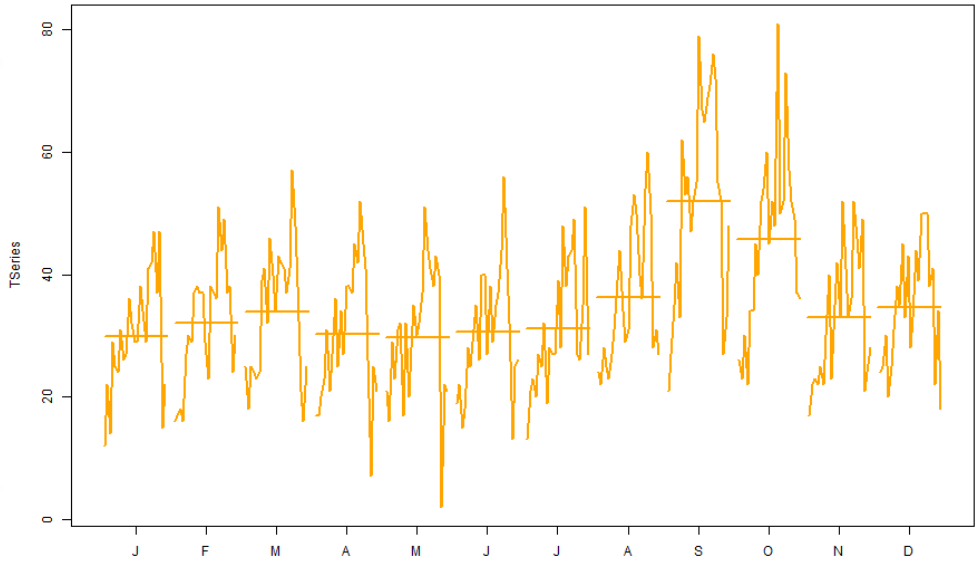


The first analysis made in order to analyse the time series is the graphical one. Our time series shows an **evident trend**, which seems to be **increasing** through years until almost **2018**. After that moment, the trend shows a **decline** and a more uncertain walk, surely due to the **Covid-19 pandemic**. The peaks seem to be relevant and to highlight a **seasonality** in the time series, which we will discuss through deeper analyses in the next slides.

Then, we decided to build a histogram in order to analyse the **distribution** of our time series data. The histogram shows some key information about the distribution of the data: the **most common monthly frequences** are between **30** and **50** movie releases. The distribution is **right-skewed**, with a **median value** of **33,5** and an interquartile range 17,75. Looking at the other statistics, we see that the **most common value** is **27**, which occurs **14 times** among the **59** different **unique values.**

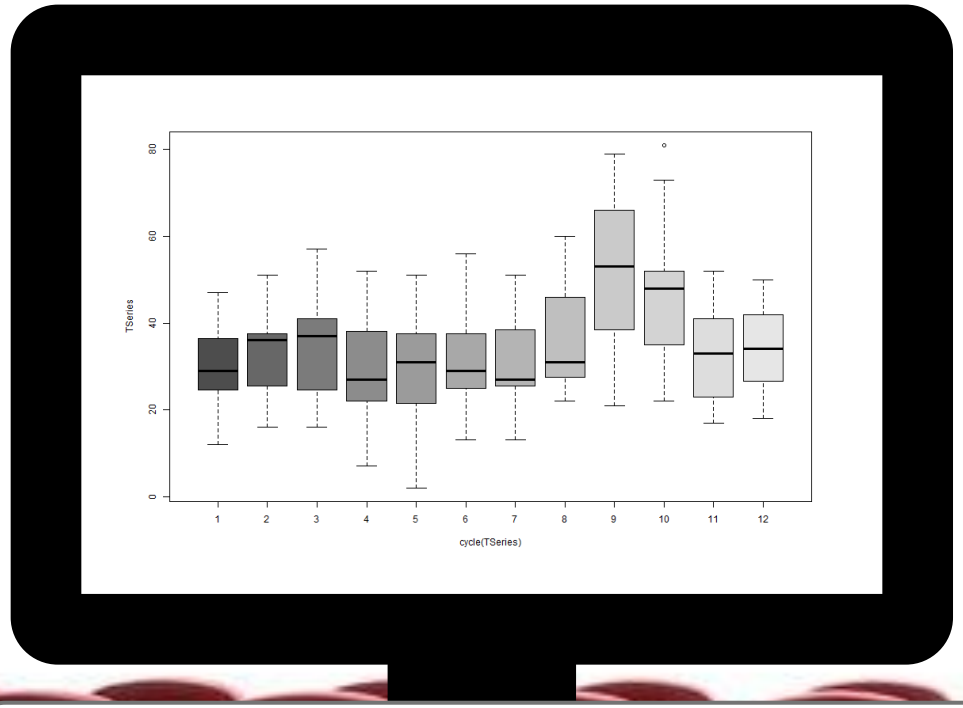
IV. Time Series Analysis – *Exploratory Graphical Analysis (2)*

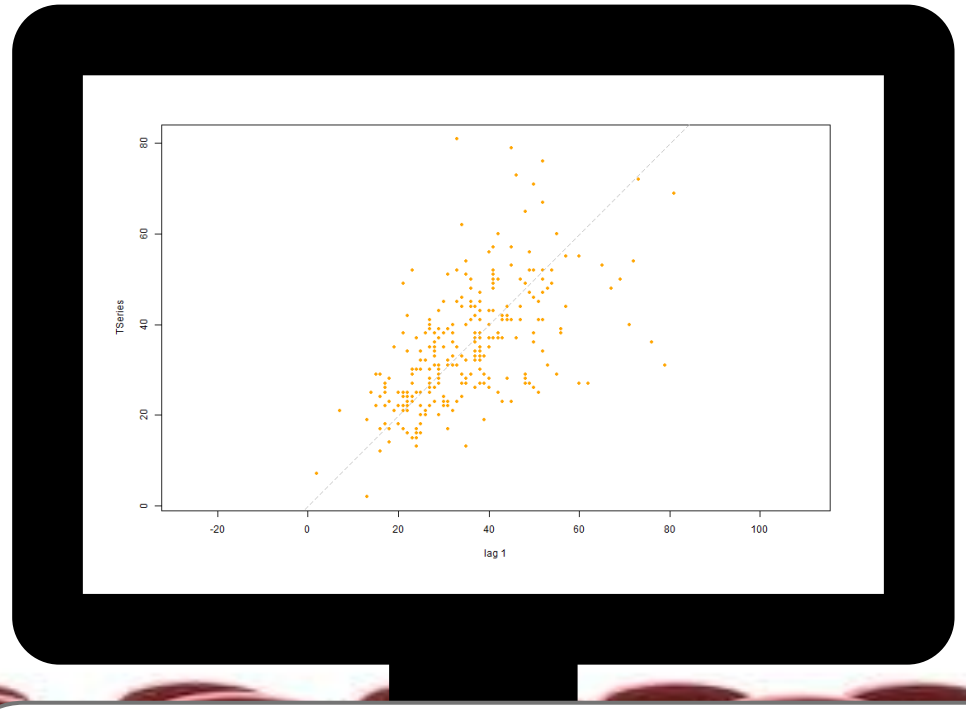




The **seasonal plot** shows that there is a **strong seasonality** in the time series over the years. In particular, the main peaks are spotted in the month of **September**. This fact highlights that the boreal post-summer period is the most active one to release movies from the production companies' point of view all over the world. We can see that the main differences in the **non-seasonal monthly data** of the time series are in **May**. In fact, if we look at the **month plot**, there is a high variability around the mean value, and in particular for the **2020** value, which is clearly affected by the pandemic effects. Moreover, with respect to the **seasonal months**, we can see two high values in **October** (the yellow and the green line), which were registered in **2014** and **2017**.

IV. Time Series Analysis – *Exploratory Graphical Analysis (3)*

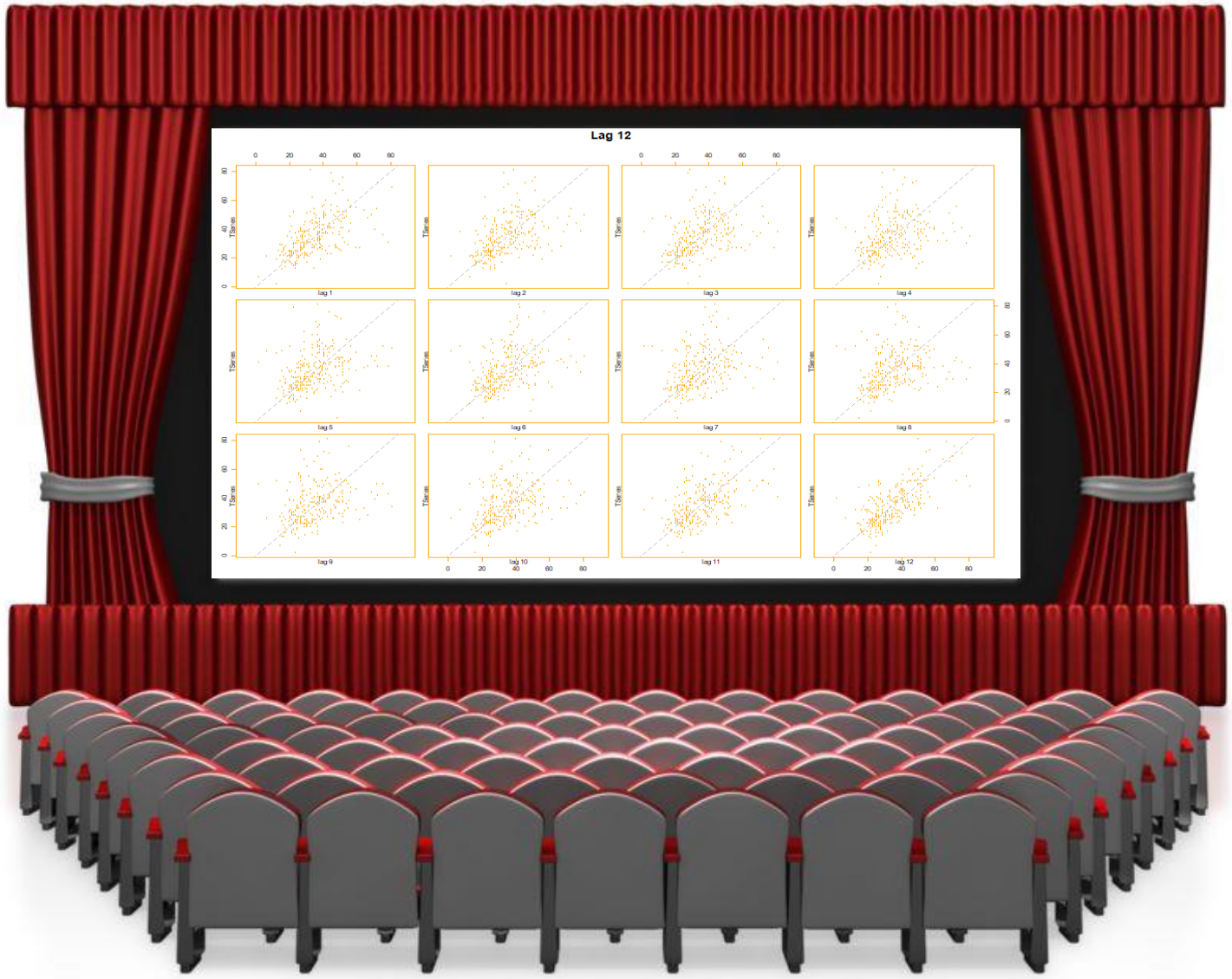




The **box plot** displays the distribution of the data over the years for each month, identifying the median value and the interquartile range for each of the 12 periods. As we saw from the analysis of the **month plot** and **seasonal plot**, the **October 2014** outlier is shown in the graphic and **significantly deviates** from the month's distribution values. Moreover, as we anticipated in the previous slide, the values of **May** are very **floating**, since they show the lowest values of the distribution, but the median value is in line with the other ones.

From the **lag plot** above, we can analyse the correlation between the monthly values of the time series with the previous month's ones. Starting from this consideration, we can affirm that the plot shows **high fluctuations** of the data and a **high variability**. As a matter of fact, the values are **not uniformely distributed** around the **line of equality,** that identifies the perfect match between each value and its previous one. This result is due to the **seasonality** of the series: in fact, the **lags of the peaks** will be the **farthest** from the line of equality.

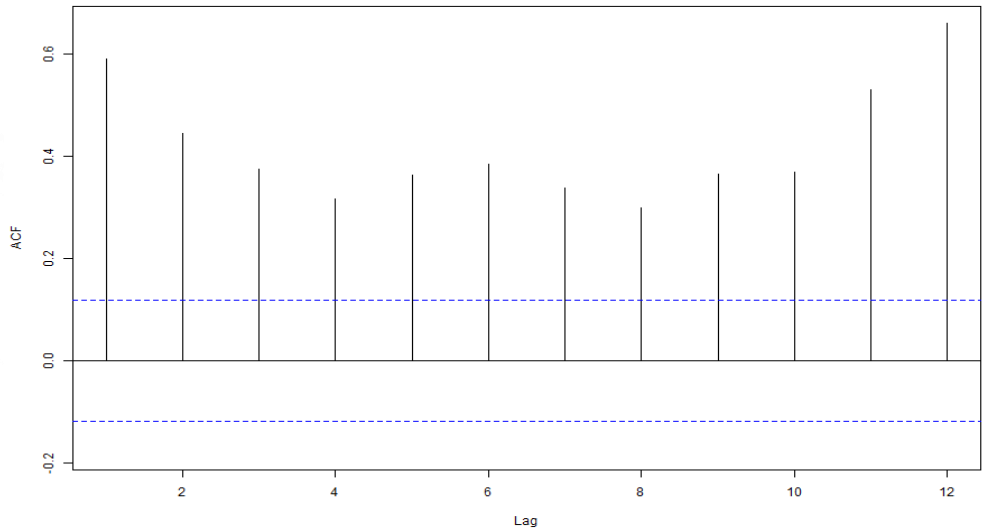
IV. Time Series Analysis – *Exploratory Graphical Analysis (4)*



We tried to investigate more thoroughly the **autocorrelation** of our time series, analysing **each lag period** in a separate plot. In fact, based on this graphic, we can more clearly determine which is the **series' seasonality**: when the points are closer to the line of equality, it means that the series' values are more closely related to one another, suggesting a possible seasonality in accordance with the studied lag. As we can see from the image, the plot with the distribution more straight to the line of equality is the **lag 12 plot**, confirming our previous consideration, stating that the seasonality of the series could be a 12 periods one (**yearly seasonal time series**).
Overall, looking at the other graphs, we can affirm that probably the **second most correlated** lag will result to be the **lag 1.** As a matter of fact, almost all the points of the plot are **close to the line of equality**. In the other cases, the distribution seems to be less concentrated around the line, and the points have a **higher variability** with respect to the previous period t-k.

IV. Time Series Analysis – *Exploratory Numerical Analysis*

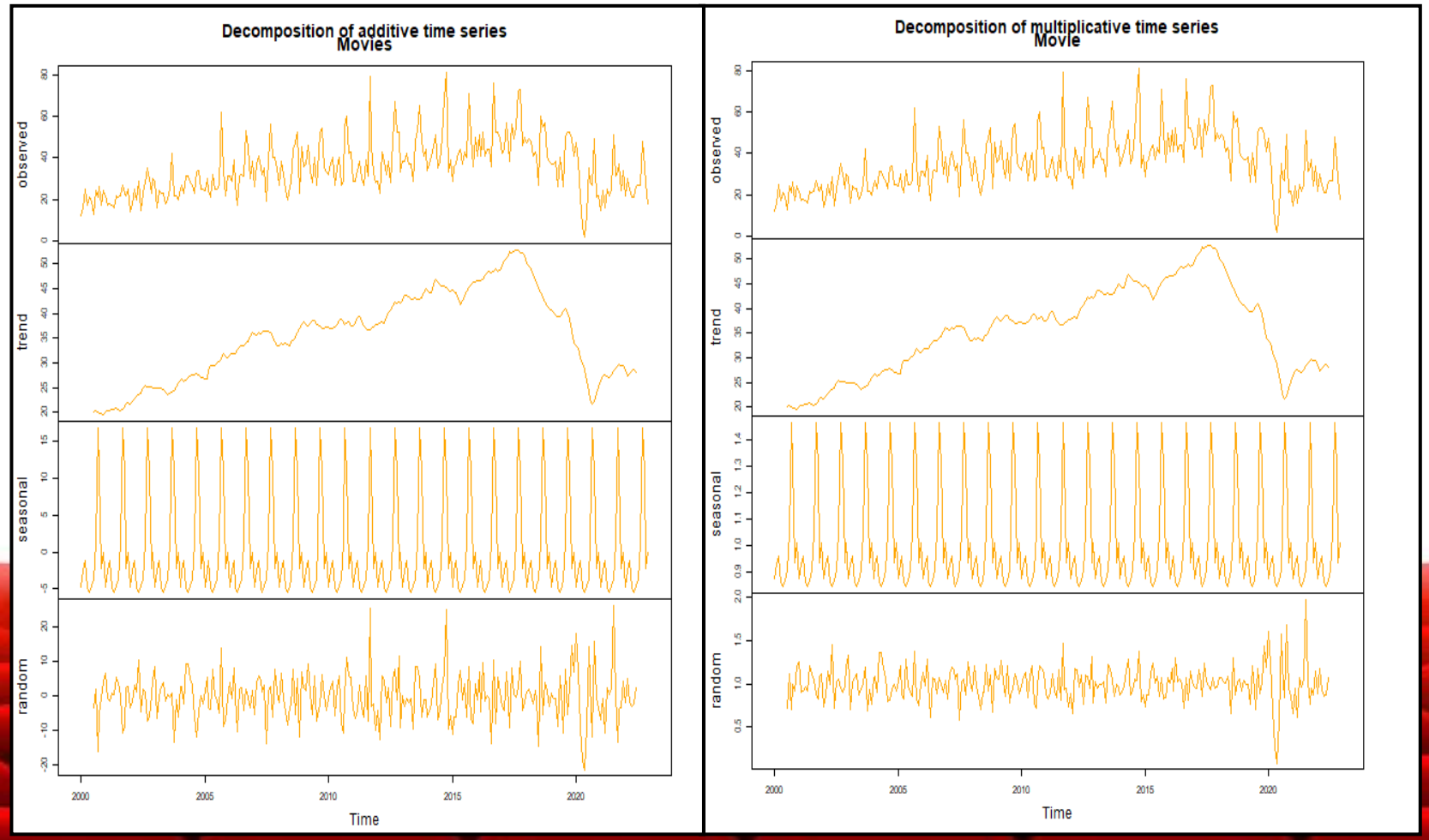
Month	Mean	Median	P95	CV
1	29,913	29	46,5	0,322
2	32,087	36	48,5	0,31
3	33,913	37	47,8	0,318
4	30,217	27	45	0,366
5	19,783	31	43,9	0,372
6	30,522	29	43,7	0,335
7	31,174	27	48,9	0,333
8	36,261	31	53,9	0,314
9	52,043	53	75,6	0,324
10	45,739	48	71,7	0,322
11	32,913	33	51,7	0,328
12	34,609	34	50	0,282



LAG	ACF
0	1
1	0,59
2	0,443
3	0,375
4	0,316
5	0,363
6	0,385
7	0,337
8	0,299
9	0,365
10	0,369
11	0,53
12	0,66

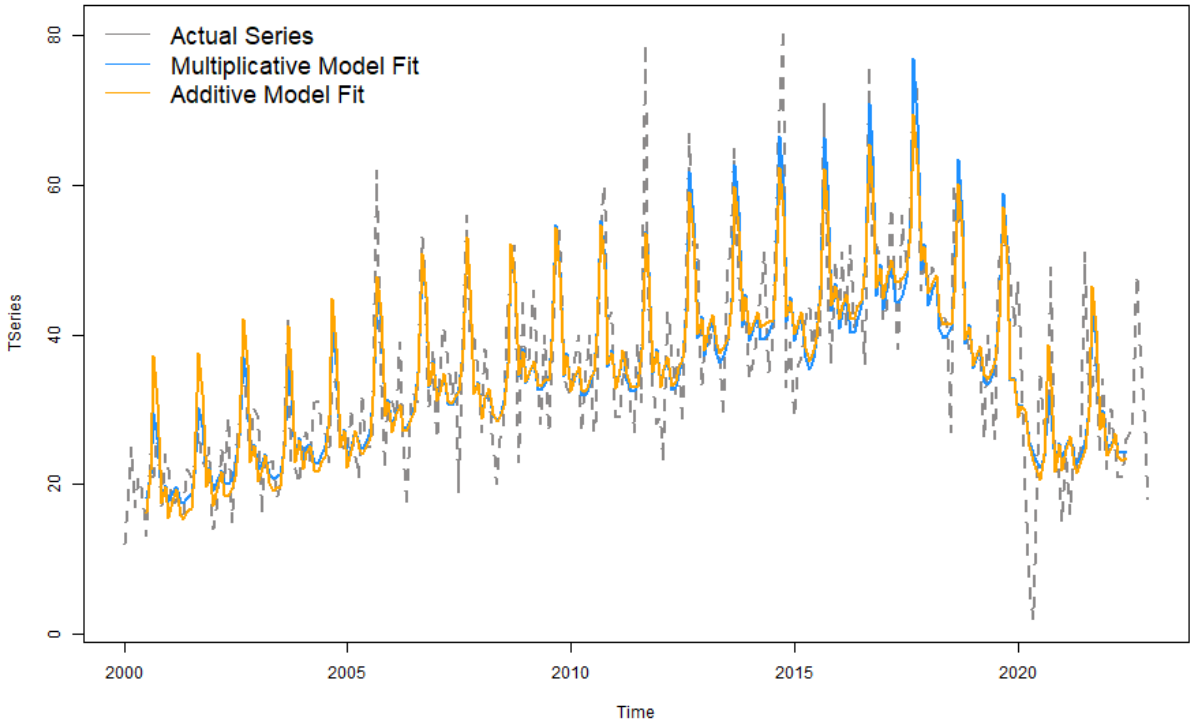
In order to go deeper inside the analysis of the time series, we made an **exploratory numerical analysis** of the **distribution** of out data and their **autocorrelation**. In the statistics table, which represents the **main statistics of our time series**, we can observe the **variability** of our data from the **coefficient of variation**. As we were considering before, the **most floating** month is **May**, which presented some lower values, in particular in the Covid-19 period. We can also observe that the **less variable** month is **December**, and the values of the **September's peaks** are **not more variable than the distribution average.** From the values of the **mean** and the **median values** we can observe the presence of a **possible right-skewed distribution** in **August** and **left-skewed ones** in **February** and **March**. Looking together at the **P95**, **median** and **mean** values, we clearly see the different peaks in **September** and **October**.
The **ACF plot** confirms the considerations we made in the previous slide, outcoming a **strong yearly seasonality** and a **high value of correlation** between the time series' values and the previous month's ones (**lag 1**). Also the **lag 11** values seem to be **strongly correlated.**

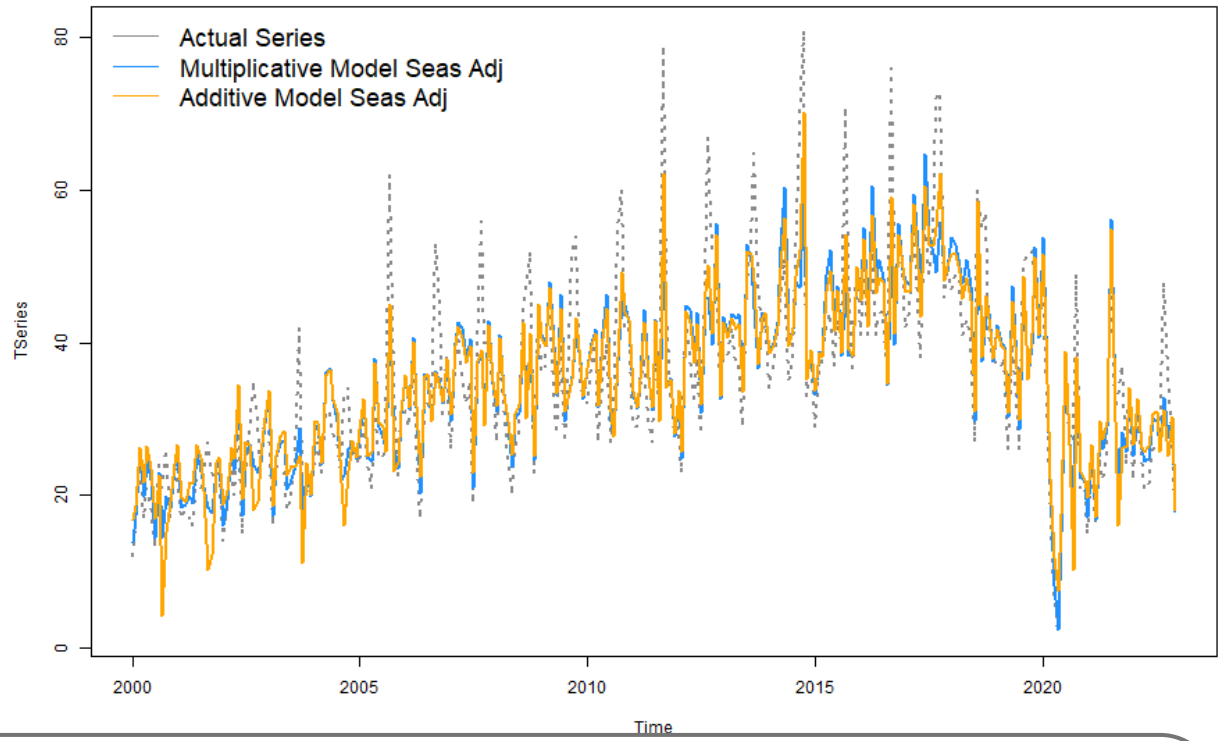
IV. Time Series Analysis – *Classical Decomposition (1)*



In order to state the **main components** of the model, we went through a **classical decomposition** of the time series. From the graphs on the left, the **trend** of the time series is evident. As we wrote before, it is **increasing** from **2000** to the **pre-Covid period,** and then starts **decreasing**. The analysis shows also the **strong seasonal component** we identified in the previous slides. Moreover, the most important feature highlighted by this analysis is about the **random component** of the model. The additive model's one shows a **less regular structure** and a **higher variability.** This is the reason why we think that the **additive model** could be **more appropriate** for our time series. In order to investigate this statement more deeply, we have developed the analysis of the **fitted values** and the **seasonal adjusted models.**

IV. Time Series Analysis – *Classical Decomposition (2)*

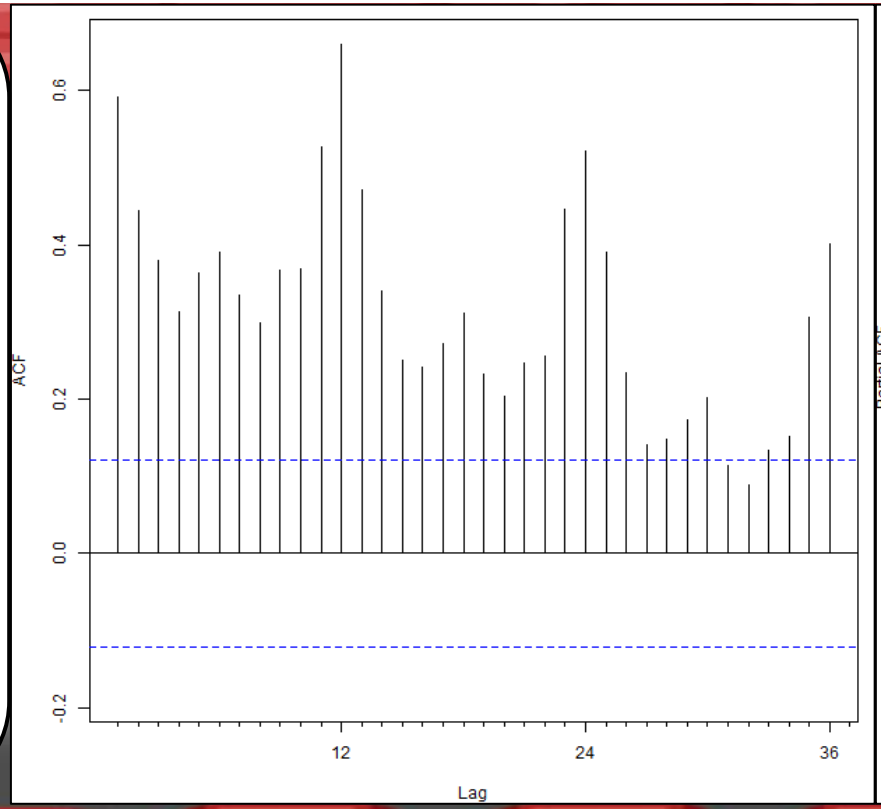


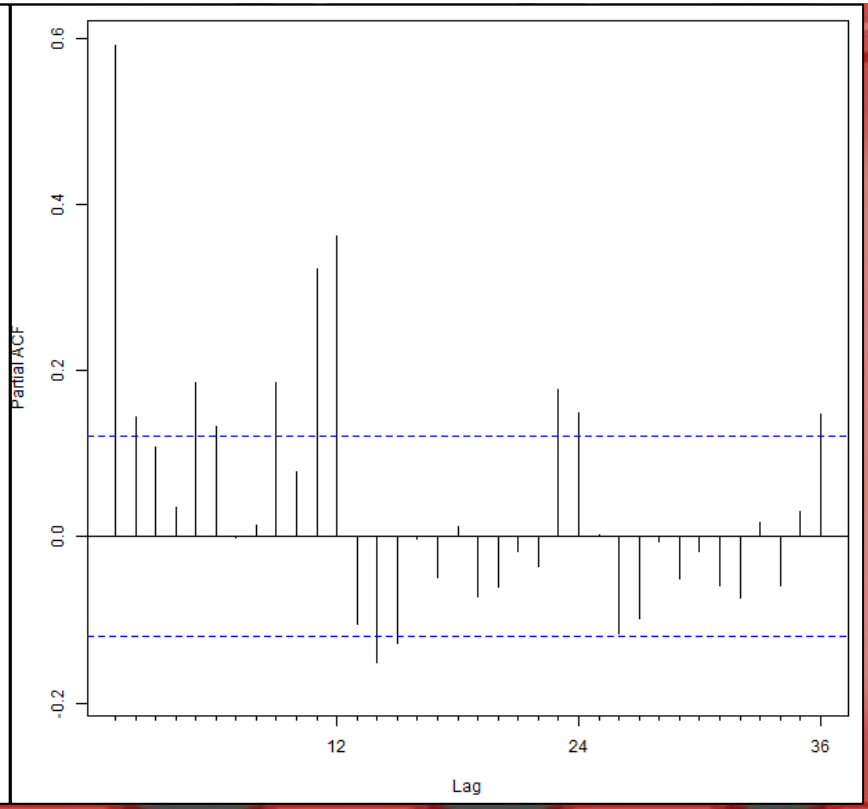


Also looking at the fit of the two models deriving from the classical decomposition, the **best** one between the **additive** and the **multiplicative** seems to be the **first one**. Actually, the difference between the two fitted values time series is a **very fine line**. In fact, if we look at the series in general, the additive model has the best fit, but if we look at **particular periods** the results appear **less obvious**. As a matter of fact, if we look at the **2012-2020** values of the model fits, we can observe that the **multiplicative model** is **more precise** with respect to the actual series. Despite this observation, the seasonal adjusted series show that the **additive model** performs **slightly better than the multiplicative one**, since there is a high amount of difference in the values between 2000 and 2012, and in the Covid phase. As a result, we can affirm **that overall the best model seems to be the additive one,** even if the best outcomes corresponding to the **trend increase** are highlighted by the multiplicative model.

IV. Time Series Analysis – *Introduction to ARIMA Models*

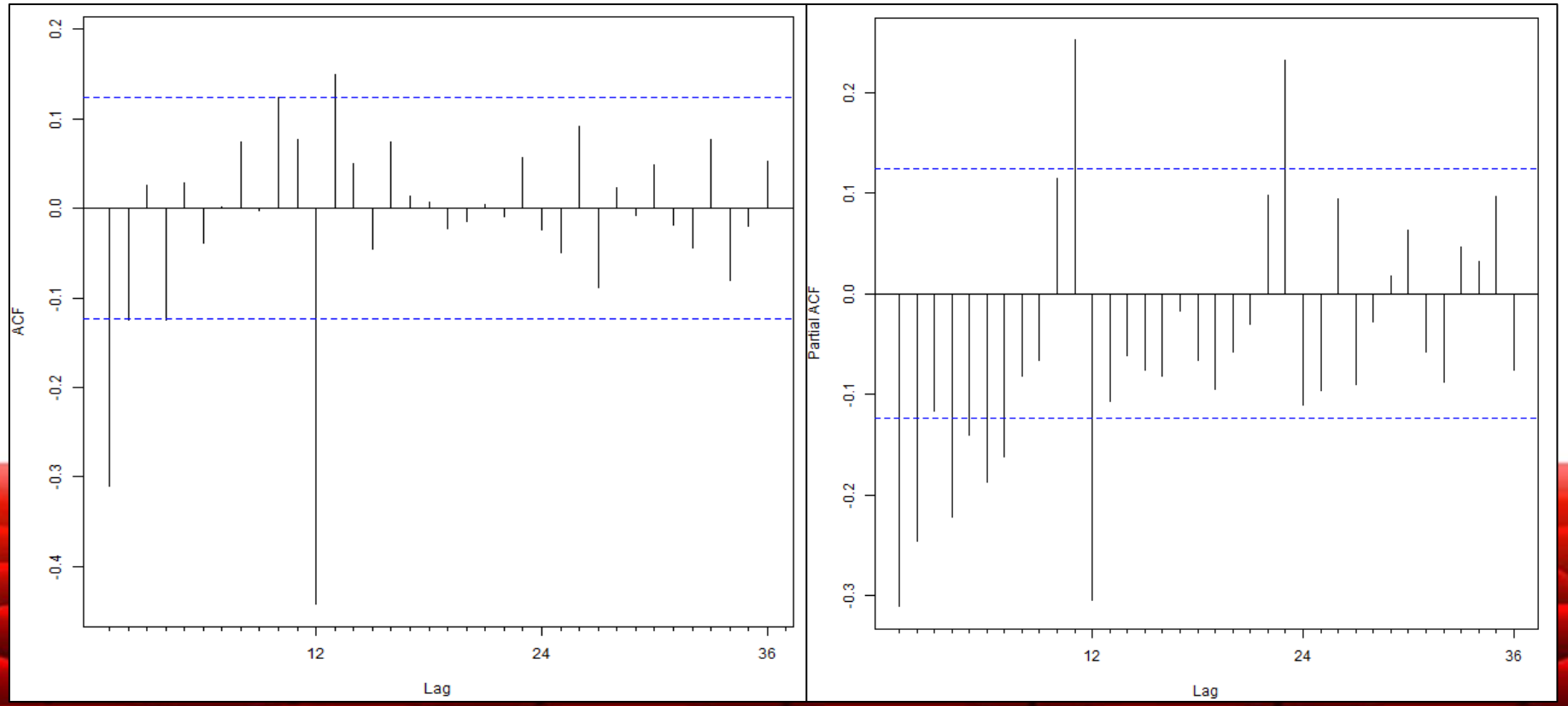
Before applying our models, we splitted the time series data into **train** and **test** sets. Since we are considering values concerning time, we split our data in a **chronological order.** We decided to keep in the **training** the period from **January 2000** to **December 2021,** evaluating **in sample** performance measures in this period. The test set is composed by the **2022** data, which are going to be used for the **out of sample comparison.**





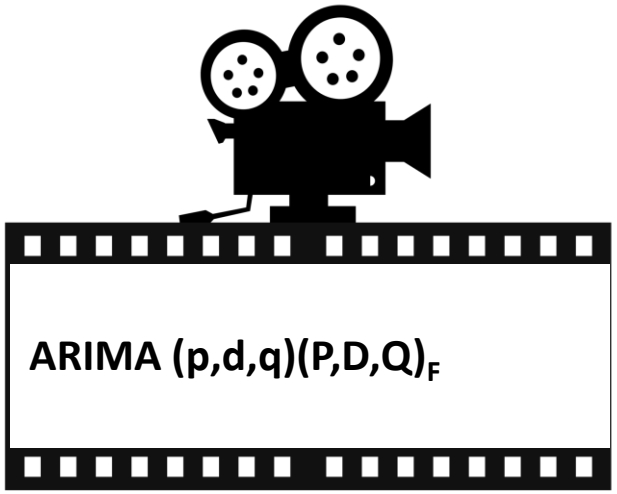
In order to predict the forecast of our time series, we tried to run different models. The ARIMA model can help us by determining the forecast of a stationary time series. Actually, as we estimated before, the time series we are considering is **not stationary**. The **ACF plot** of the **original series** shows the **gradual decrease** in the **autocorrelation values,** reaching the peaks every 12 months. There are both strong seasonal and trend components. The **PACF plot**, on the other hand, shows the correlation between the time series and its lagged values after removing the effects of intermediate lags. Looking at this plot, we can state **that almost each seasonal peak of the original series is significant.** In order to reduce the trend and induce stationarity in the time series, we tried to make a **differencing** and **a log transformation** of the considered time series.

IV. Time Series Analysis – *ARIMA Models: Introducing Stationarity*



In order to induce **stationarity** in our time series, we tried to make a **differencing** and **a log tranformation**. The ACF and PACF plots on the left show that the **significant peaks of autocorrelation** are **less frequent than before**. As a result, we can affirm that we **stationarized the series successfully**. To perceive our main objective, which is to state which are the best ARIMA models that fit out time series, we looked at **these two plots** trying to build **8 different models** with the **6 different parameters.**

IV. Time Series Analysis – *ARIMA Models: Parameters Estimation*



N_DIFF	N_SEAS_DIFF
1	0

Row ID	aic_corrected	p	d	q	P	D	Q	F
arima_1	113,882	0	1	1	1	0	1	12
arima_2	**96,306**	**1**	**1**	**1**	**1**	**0**	**1**	**12**
arima_3	**98,274**	**1**	**1**	**1**	**2**	**0**	**1**	**12**
arima_4	115,949	0	1	1	2	0	1	12
arima_5	100,351	1	1	2	2	0	1	12
arima_6	**98,375**	**1**	**1**	**2**	**1**	**0**	**1**	**12**
arima_7	**99,957**	**0**	**1**	**2**	**2**	**0**	**1**	**12**
arima_8	**98,004**	**0**	**1**	**2**	**1**	**0**	**1**	**12**
arima_auto	109,488	1	1	1	2	0	0	12

Starting from the previous **ACF** and **PACF** plots, we tried to assume **different ARIMA configurations**. We built these **8 models** making some hypotetical assumptions about the values of the **non-seasonal parameters (p, d, q)** and the **seasonal** ones (P, D, Q). With respect to the **d** and the **D** components, we ran the «ndiffs» and the «nsdiffs» functions in R. These functions state the **non-seasonal** and the **seasonal differences** of the time series that we should consider, that were respectively 1 and 0. After that, we looked at the **PACF** plot to understand the possible values of **p** and **P**, which represent the **autoregressive** terms of the model. We estimated that the value of p could be 0 or 1, and the value of P could correspond to **one** or **two seasonal** years. Looking at the **ACF** plot we made some assumptions about the parameters **q** and **Q**, suggesting the right **moving average** components. The value of q seemed to be 1 or 2, while the seasonal component Q was identified as one seasonality. As we can see from the table, the values stated by the **automatic function** in R do **not have any correspondence** with the models built by us.
The corrected value of the Akaike Information Criterion (AICc) is a goodness of fit measure considering the number of estimated parameters and the sample size of the model. We used this metric in order to understand which one of the stated ARIMA models could be the best one in the analysis. From this analysis, the best model seems to be the «**arima_2**», which shows the **lower AICc** value. In order to go deeplier into che analysis, we stated other error measures based on the in sample comparison.

IV. Time Series Analysis – *ARIMA models: best model estimation*

Row ID	ME	RMSE	MAE	MPE	MAPE	MASE
Arima_1_Test	**-0,293**	**4,675**	**3,165**	**-4,528**	**12,229**	**0,397**
Arima_2_Test	0,792	5,041	3,428	-0,692	12,609	0,43
Arima_3_Test	0,533	4,981	3,321	-1,731	12,319	0,416
Arima_4_Test	**-0,169**	**4,685**	**3,162**	**-4,04**	**12,175**	**0,396**
Arima_5_Test	0,463	4,952	3,305	-2,001	12,3	0,414
Arima_6_Test	0,778	5,034	3,427	-0,751	12,62	0,43
Arima_7_Test	**-0,169**	**4,685**	**3,162**	**-4,04**	**12,175**	**0,396**
Arima_8_Test	0,463	4,952	3,305	-2,001	12,3	0,414
Arima Auto_Te	3,331	7,912	6,037	8,582	21,55	0,757

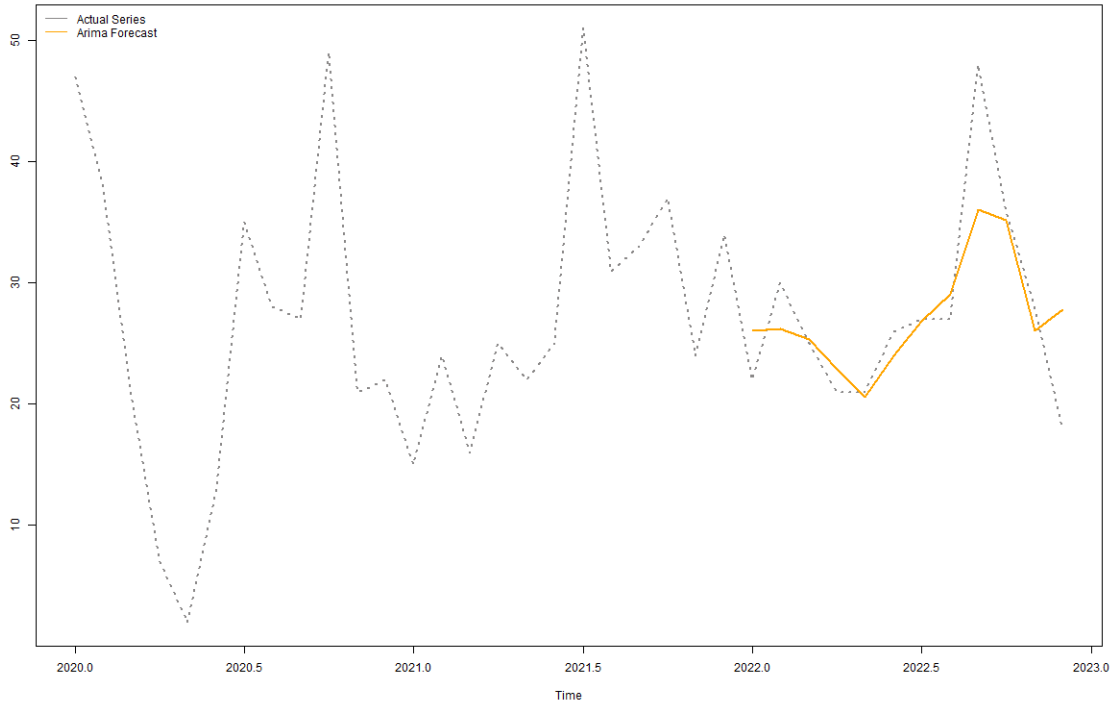
In order to estimate the **best ARIMA model** to keep and compare it to the other models we are going to build, we ran an **in sample analysis** (on the training data). We used different **accuracy**, **scale-dependent** and **bias measures** to understand our results. The first accuracy measure we looked at was the **MAPE**, which shows the mean absolute percentage error, and permits to compare different models. As we can see from the table on the left, we have **two lower values**, which correspond to the **Arima_4** and to the **Arima_7** models. Also the **MASE** value is the best for these two models show better fits than the average naïve forecast computed on the training data. Also the bias measures state that these two models are the best for the forecasting. In particular, the **MPE** shows that the models **overforecast data by 4.04%.**

Since the two models showed the same model estimation values looking at these accuracy, scale-dependent and bias measures, we decided to state the best model looking back at the **lower AICc** value between these two models. The **Arima_7** AICc is 99.957, stating that this is the best model of our in sample analysis, corresponding to the **ARIMA (0,1,2) $(2,0,1)_{12}$**

IV. Time Series Analysis – *ARIMA models: out of sample best model estimation*

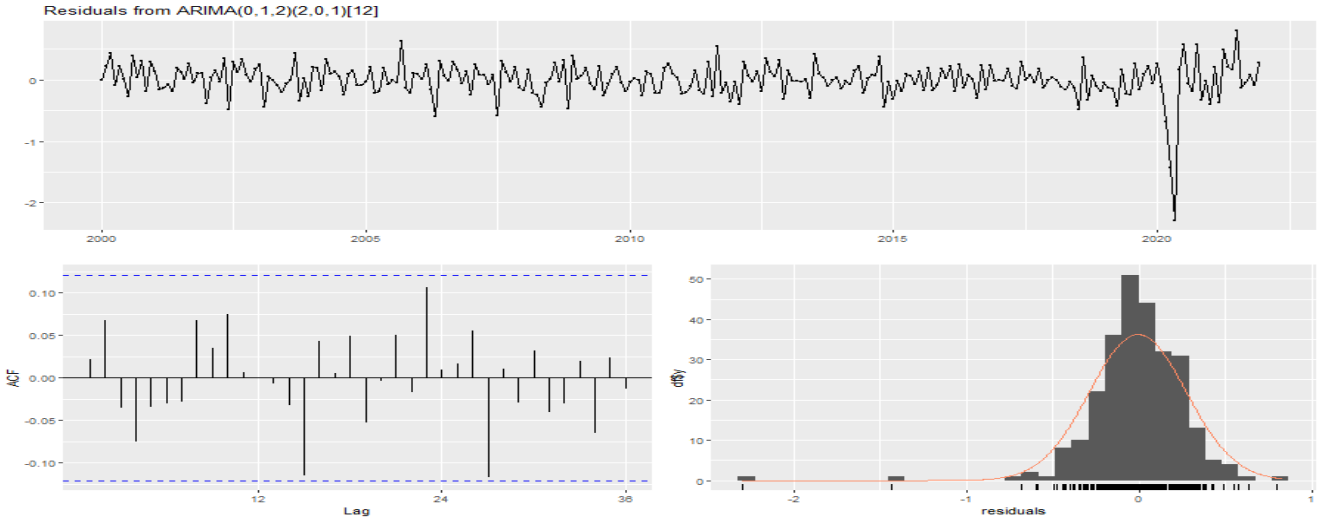
Row ID	ME	RMSE	MAE	MPE	MAPE	MASE
Arima Test	**0,244**	**4,879**	**3,266**	**-2,817**	**12,309**	**0,409**
Arima SIM 1	-6,195	8,853	7,391	-19,199	21,626	1,025
Arima SIM 2	**2,75**	**6,375**	**4,644**	**3,877**	**8,567**	**0,667**

	Train Set	Test Set
SIM 1	2000-2018	2019
SIM 2	2000-2016	2017



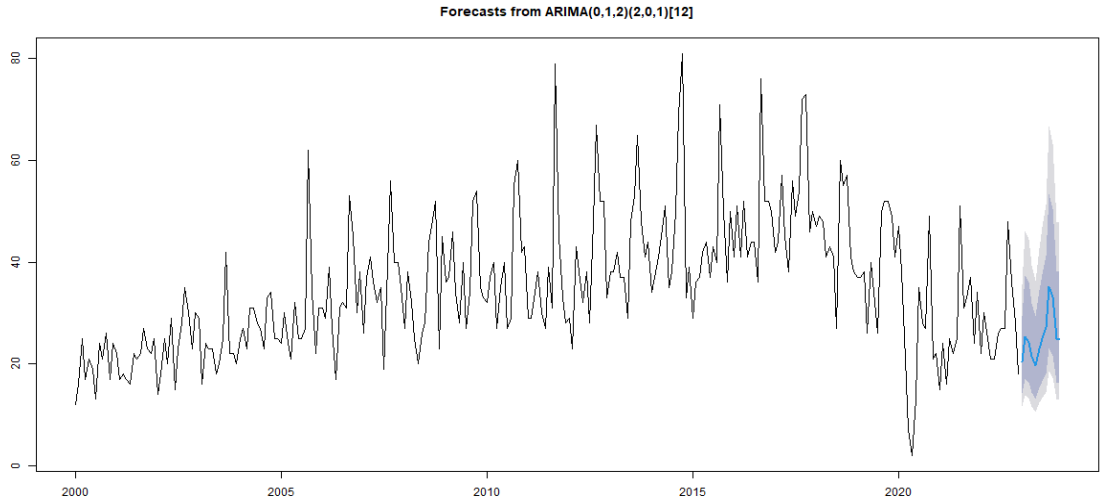
We used the **Arima_7** in order to make a **forecasting** on the training set and make an **out of sample analysis**. We also run two other simulations using different train and test sets in order to understand how the model fits for other historical periods. The first simulation (**SIM 1**) was made in order to predict the time series values of the **last pre-pandemic year.** The second one (**SIM 2**) had the aim to understand if the very high values that we saw in the graphical analysis for the **October 2017** had an **impact** on the model, or if they were predictable. However, the **simulation 2** measures showed the best results in terms of **MAPE**, which means that those values were in line with the model. On the other hand, the best **MASE** result was obtained by our «general» model. As a result, **our best Arima model overforecasts data by 2,817%.** It also provides, on average, better forecasts than the average naïve forecast computed on the training data.

IV. Time Series Analysis – *ARIMA models: 12 months «real» forecasting*



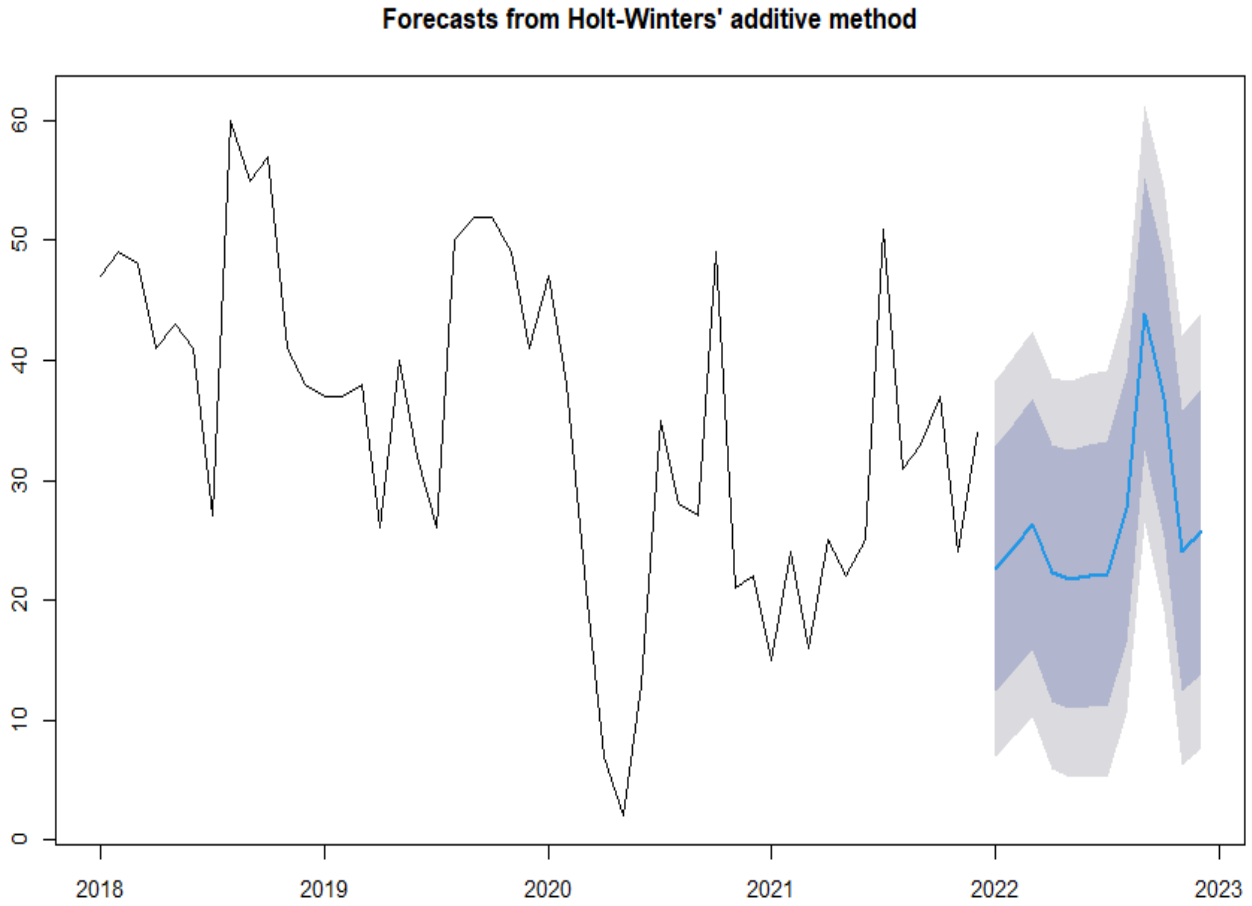

```
data:  Residuals from ARIMA(0,1,2)(2,0,1)[12]
Q* = 17.318, df = 19, p-value = 0.5683
```

The last statement about the model was made looking at the **residuals** . We can easily see that the **ACF plot** shows **randomical** values, so there is **no autocorrelation** between the residuals in our model. The **distribution** of the residuals is **normal**, and as we can see from the image above the **p-value** shown by the R script evaluation is **not significant.**



In order to complete our ARIMA model analysis we tried to make a **«real» forecasting** on a period that we have not considered in our time series. As a matter of fact, we forecasted the **2023** values with the forecast function in R, which estimated that the 2023 values could be **highly influenced by the pandemic period.** In fact, the forecasted **peak** seems to be very **low** (blue line), with a **high possible error** (grey space). This high error value are probably due to the **instability** of the previous years, where there have been both **increasing** and **decreasing trends**.

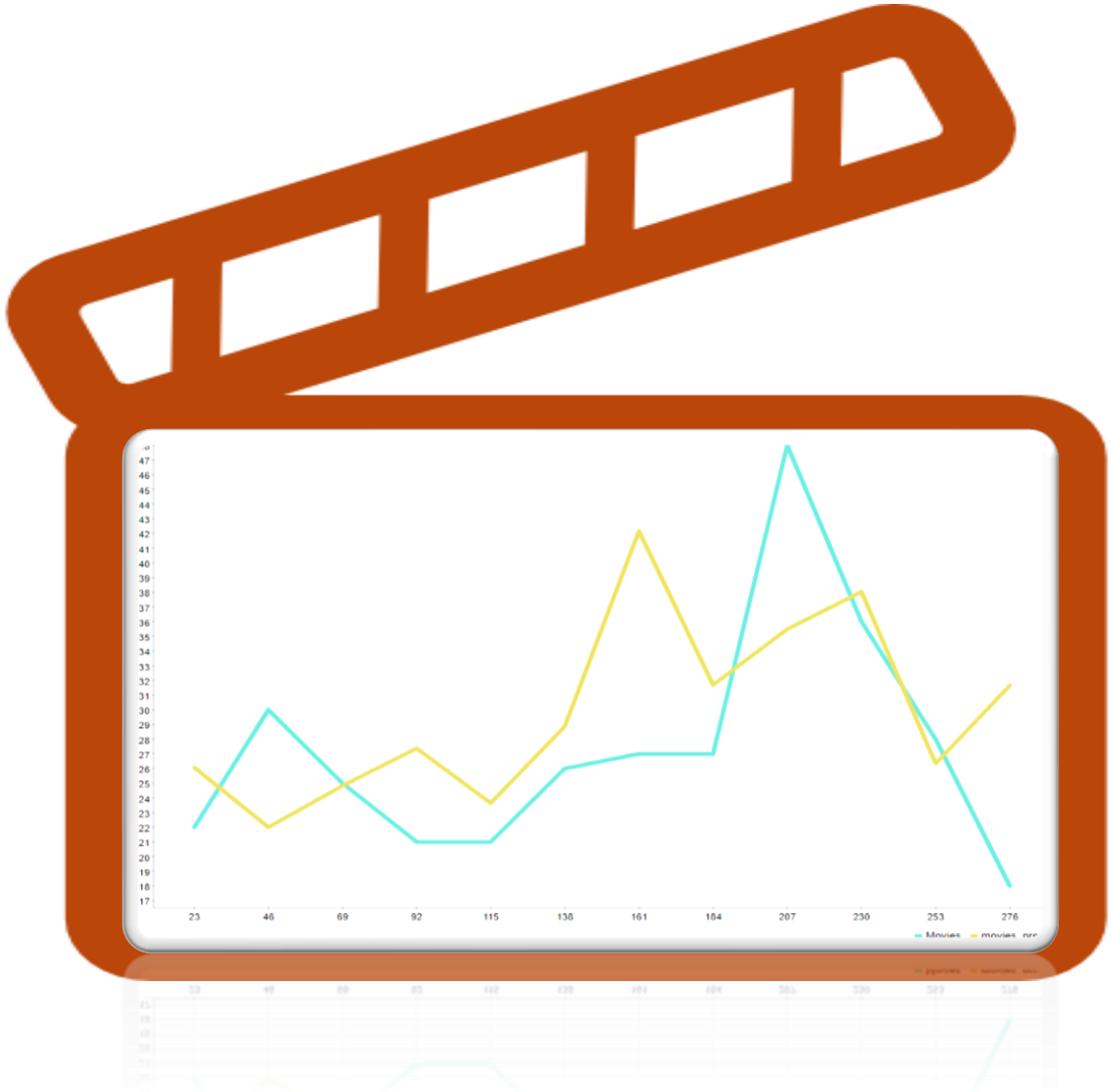
IV. Time Series Analysis – *Exponential Smoothing*



Row ID	ME	RMSE	MAE	MPE	MAPE	MASE
HW Add Test	**0,754**	**3,72**	**2,953**	**0,689**	**11,6**	**0,37**
HW Mult Test	1,173	4,44	3,573	0,996	13,078	0,448
Auto ETS Test	-1,804	4,497	3,598	-10,214	15,255	0,451
SES Test	-2,681	8,161	6,648	-16,988	25,937	0,833
Holt Test	-3,22	8,319	6,965	-19,064	27,47	0,873

Another useful method to make forecasts of our time series is the **exponential smoothing**. We decided to investigate the fit of the time series using different **ETS models**. We expect the **Holt-Winters' models** to be more **performing** because we know that they are more appropriate for time series with a clear **trend** and **seasonality**. Also in this case, we tried to run both the **additive** and the **multiplicative** Holt-Winters' methods, and as a result we obtained that the **best** one is the **additive** one. In fact, the **out of sample comparison** measures of the different models show that **all** the considered **measures** are **better** for the Holt-Winters' additive method. The **MPE** shows that the model **slightly underforecasts** data by **0.689%.** The **MASE** is **lower** than **one** and the lowest among the models taken under consideration.

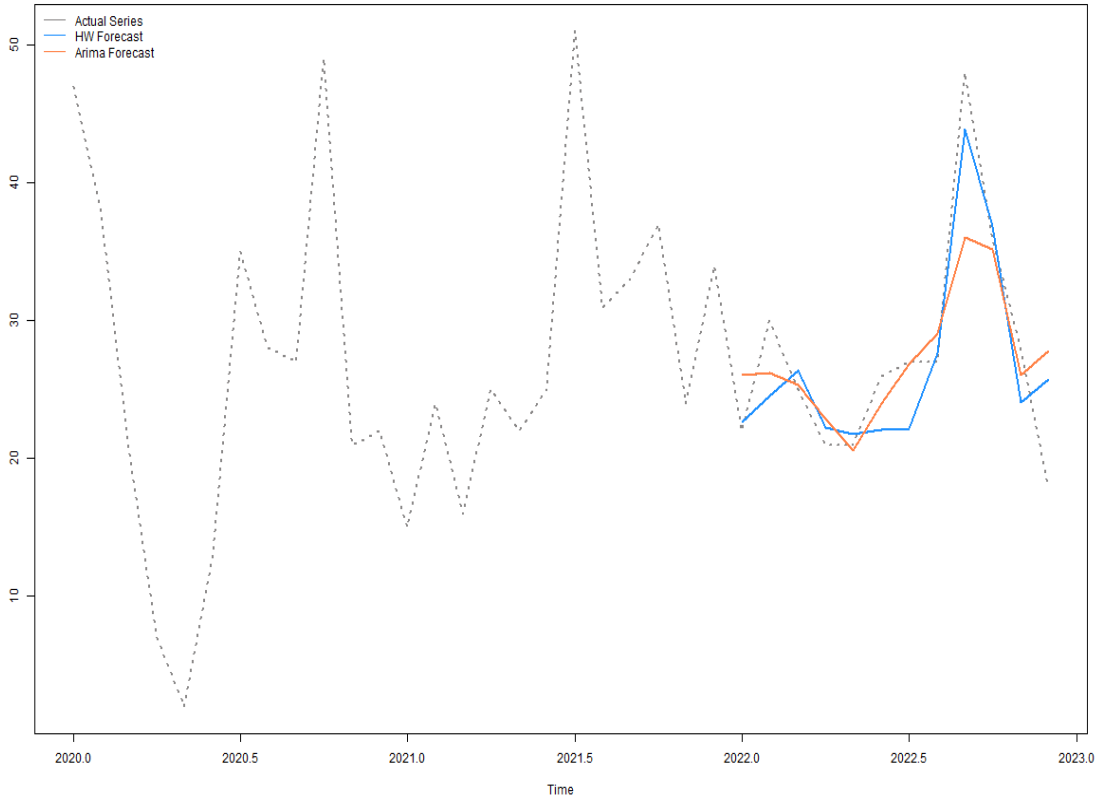
IV. Time Series Analysis – *MLP Forecast*



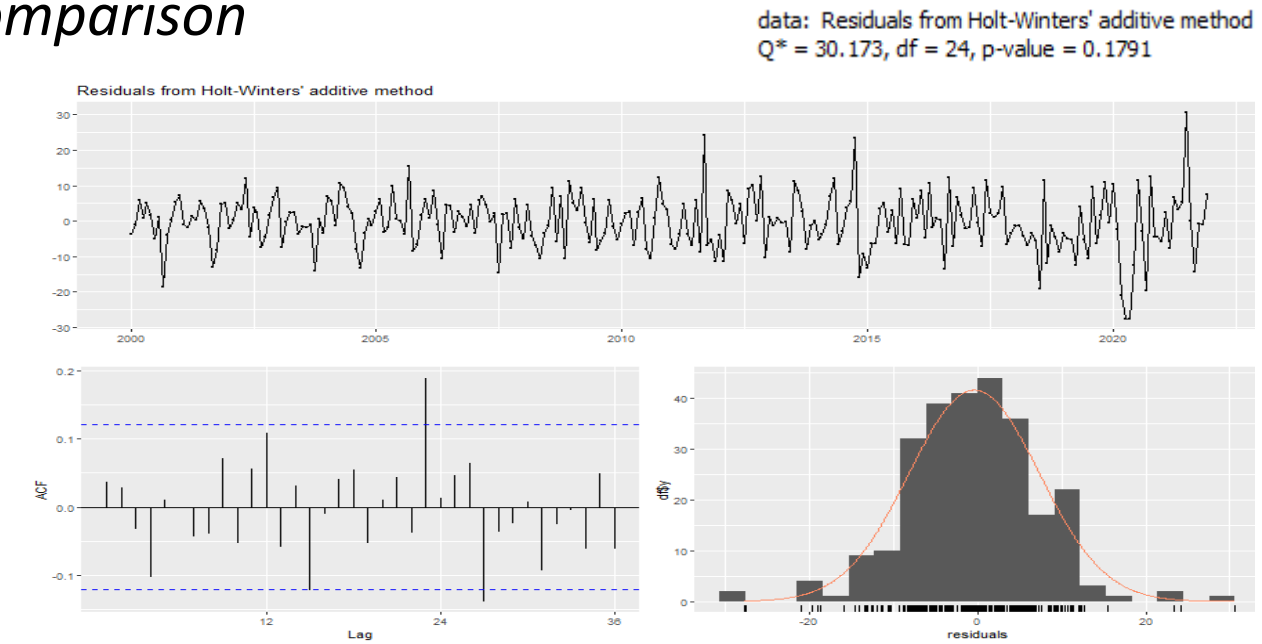
We tried to run a **multilayer perceptron** model to evaluate our time series. Since our series could have some **outliers** due to the **Covid-19 pandemic**, we tried to use the «**tsclean**» function in order to clean the time series from outliers and abnormal values. We built the model in different steps:

1. Creation of **12 different lag columns** for each value starting from **2001** (the 2000 data would have missing values in the columns of the first year)
2. **Partitioning**, using the period 2001-2021 as train set and 2022 as test set.
3. **Normalization** of all the values of the lags between **0** and **1**
4. **Fitting** of a **MLP model** with **one hidden layer** and **11 hidden neurons** through the «RProp MLP Learner» node.
5. Out of sample **prediction** through the «MultiLayerPerceptron Predictor» node
6. **Denormalization** of the predicted values
7. **Computation of MAPE,** which is equal to **0.239**

IV. Time Series Analysis – *Models comparison*



Row ID	ME	RMSE	MAE	MPE	MAPE	MASE
Arima SIM 2	**2,75**	**6,375**	**4,644**	**3,877**	**8,567**	**0,667**
HW SIM 2	2,22	6,18	4,797	2,643	8,94	0,689
HW Test	**0,754**	**3,72**	**2,953**	**0,689**	**11,6**	**0,37**
Arima Test	0,244	4,879	3,266	-2,817	12,309	0,409
Arima SIM 1	-6,195	8,853	7,391	-19,199	21,626	1,025
HW SIM 1	-5,9	8,777	7,441	-18,531	21,654	1,032



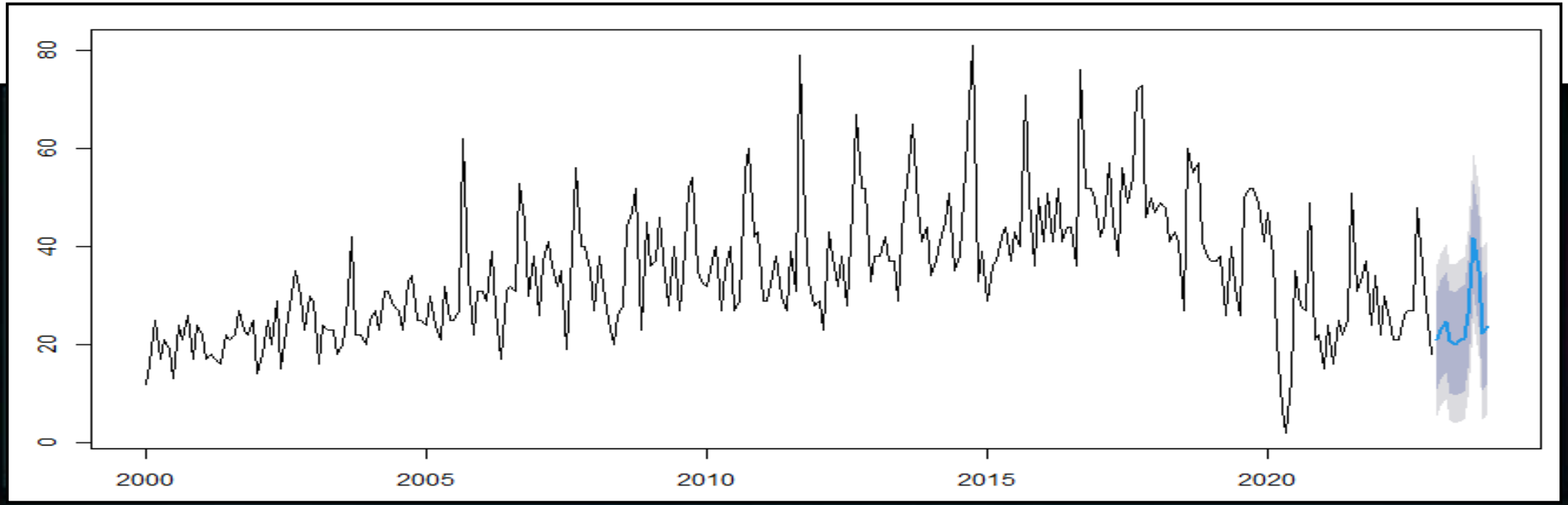
Finally, we compared the two best models we found from the previous analysis:

1. **ARIMA (0,1,2) $(2,0,1)_{12}$**
2. **Additive Holt-Winters' exponential smoothing**

As a result, the **best model** for the training and test we want to consider is the **Holt-Winters' ETS** one, but if we change the partitioning, the situation will change significantly. As a amatter of fact, we obtain the best **MAPE** performances in the **Simulation 2 Arima** model.

On top, we looked ad the **ETS model residuals**, which **do not seem autocorrelate**, are **normally distributed** and assume a **non-significant trend,** as we can see from the **high p-value.**

IV. Time Series Analysis – *Conclusions*

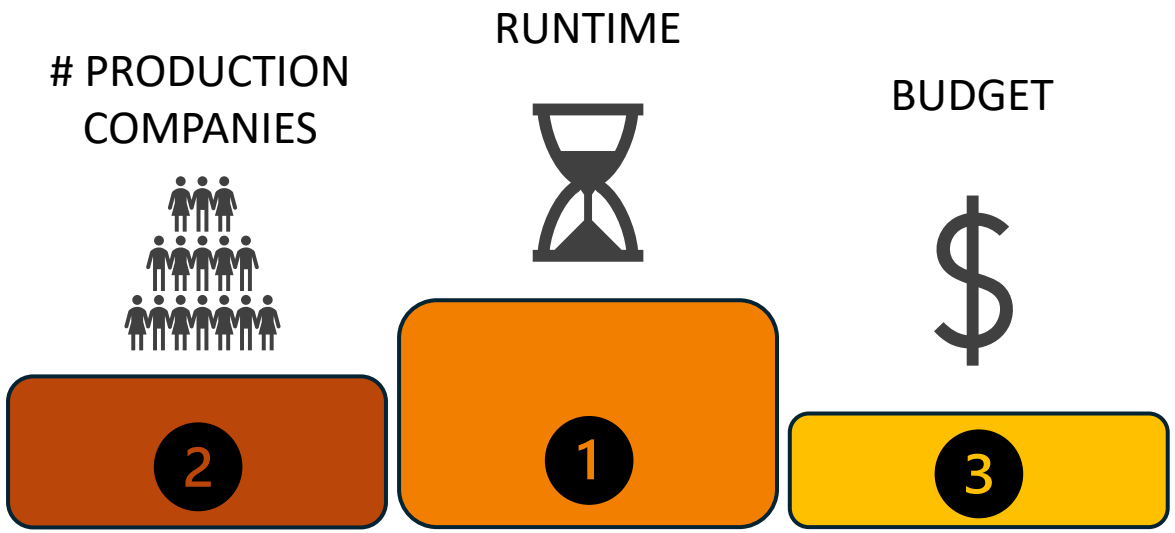


In the time plot above we can see the **forecast** made with the **additional Holt-Winters' exponential smoothing method**, which resulted to be the best among all the models built for our time series analysis. From the image, we can see that the **effect** of **Covid** is **determinant** in order to make a prediction of the **2023** values. In fact, the **trend** seems to be **smooth** and the **2023 peak** is **lower** than the previous ones, since a **very high instability** was registered in the years before. We expected the 2023 results to be better because of the almost complete recovery from the pandemic situation, which could have had some notable effects until 2022 due to the scheduling phase of the movies. Despite that, the Hollywood strike could have had another strong exogenous impact on the time series, since the US data are the majority of our sample. Overall, the series considers only the Covid-19 shock, and this is the reason why the seasonal peak is not high as the others.


MANAGERIAL
IMPLICATIONS

V. Managerial Implications – *Best models implications*

As highlighted by the Random Forest model, the variables that have the greatest impact on the success of a movie are:



There are no standout variables; rather, there are **other important variables**. Additionally, the **number of genres** represented in the movie, the **country of production**, and **the original language** being **English** are crucial. These latter two variables serve as a proxy for the ongoing significance of American know-how in a movie's success.

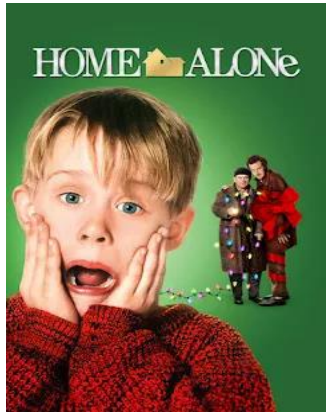
For a qualitative interpretation of these variables, whether they **positively** or **negatively** impact, we rely on the **Logit model**. By **cross-referencing** these models, we can provide higher-quality interpretations of our results. For now, we can only say that The Space should pay attention to these variables when they must consider to buy a movie.

V. Managerial Implications – *Movie Genres*

While performing bivariate analyses we found out an interesting pattern according to which those genres that have, on average, the **highest budgets** (**Adventure**, **Animation**, **Sci-fi**, **Fantasy**, **Family**) also had the **highest revenues**.
On the other hand, ***Documentaries*** and **TV Movies**, which were shown having the **lowest budgets**, also had the **lowest revenues**.

By comparing this trend with the insights on movie genres from the logistic regression, where **Animation**, **Sci-fi**, **Fantasy** movies showed odds-ratio > 1, while **Documentaries** and **Family movies** had an odds-ratio < 1, we notice that:

Animation movies have, on average, high budgets and high revenues.
At the same time, this genre is positively associated with success (the odds of a movie being successful are higher when the movie belong to animation genre compared to when it doesn't belong to it).



Family movies also have, on average, high budgets and high revenues.
But, looking at the logistic regression, the odds of a movie being successful are lower when the movie belong to family genre, ceteris paribus.

First Take-away

TheSpace should consider investing more in acquiring rights to **animation movies**. This targeted approach aligns with current market trends and audience demand for animation content**.**

Second Take-away

The trend discrepancy in family movies can be explained by the typical viewing behavior of the target audience. Perhaps family members are usually causal viewers who perceive watching movies as **social gathering occasions** and do not really care about the content itself. **TheSpace** can consider **themed events** and **interactive activities** to **enhance the customer experience**.

V. Managerial Implications – *Runtime and Number of companies*

Runtime

As we were considering in the previous slides, «**runtime»** is the **most important variable** to consider in order to **predict** our **target variable**. The high predictive power of this variable is highlighted by the **random forest model**, but it was also previously suggested by the **bivariate analyses**. As a matter of fact, we noticed a **significant correlation coefficient** of **0.254** between the runtime of a movie and its average vote on the online platform. Also the **Gauss Logit** shows a high positive coefficient for the runtime variable, which confirms the positive association with our target variable, which tries to approximate the popularity of the movie.

As a cinema, the best way to take advantage from this important insight could be to **maximize** the revenues considering that **a movie with a long runtime could bring higher incomes** even if it could have a **negative impact on the halls turnover.** For example, the incredible success of Oppenheimer in 2023 can confirm this analysis, since the movie has a runtime of **180** minutes, while the average is around **110** minutes.

Number of companies

The variable «**Number of companies**» results to be the second most important variable in the random forest. In the bivariate analyses we saw that number of companies has also a **high correlation** with the variable **«Number of countries».** The value of this correlation, which is **strongly positive** (0.56 with a significant p-value), suggests that movies produced by more than one company are often associated with an **international involvement**. Moreover, we can see that the number of companies impacts positively on the target variable, even if the odds ratio is not the highest of the model. This fact, together with the impact that the variable «Number of companies» shows on our target variable, leads us to investigate more on the **specific companies** that have an impact on our model. Looking at the Gauss Logit, we notice that the **production company** with the **higher impact** on the target variable is **Walt Disney.** If we try to combine all these considerations, we will come up with some managerial implications for our cinema. In fact, a **multi-company produced movie** could be the best investment in order to project popular movies. Moreover, the **involvement** of **different countries** in the **movie registration** could be a good performance indicator.

V. Managerial Implications – *Time series implications*

Our time series shows some **important features** which could **imply** very useful **insights** about our **cinema management**.

Employees management

We can see different **peaks** in the number of movies released in the **different months** of the year. It means that, when a **higher number** of **movies** will be **released** in the same period, the **saturation** of the halls will be **higher**. Probably a **higher number of films** will be **projected** in these periods, which according to our analysis will be **September** and **October**. This is the reason why the **number of employees** working in the cinema should be **higher**, in order to **satisfy** larger amounts of **crowd** in the **rush hours.**

New customers acquisition

All the periods of the year when the number of published movies decreases is an opportunity to implement discounts and special offers. In fact, if people will be more encouraged to go to the cinema when the number of projected movies is lower, they will have a higher willingness to pay in the seasonal peaks.

Halls turnover

During the **most intense** periods of film releases, there are more movies that could be projected in the **cinema halls**. This is the reason why in specific periods the **turnover** of the halls should be very **fast**, in order to **maximize revenues**. In this way, the number of **available slot** per day could **increase**.

Limitations

At the end of the project, we can highlight additional limitations that could have improved the model.

Target Variable

Due to the nature of the dataset, the lack of many pieces of information, and to avoid losing too much data to the extent that we couldn't implement our model, we had to construct a target variable with logged and standardized variables.

Number of N/A

If the dataset had all the data available and not recoded as zero within the original dataset, we could have built models based on 1 million observations. In this case, data cleaning reduced the number of observations by 99%, but on the other hand, it would not have been possible to continue the analysis otherwise.

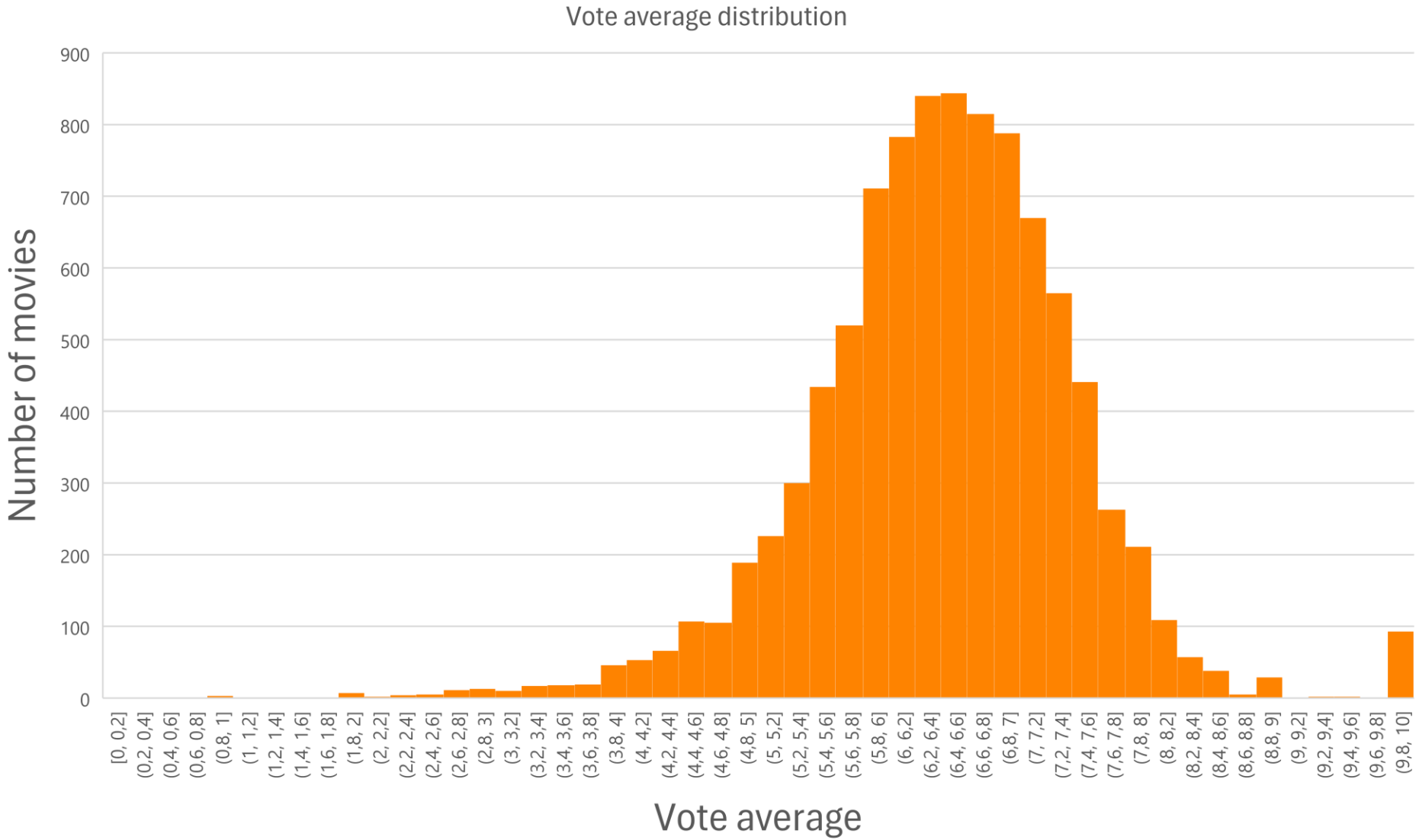
Many values have been logged

Due to the nature of the data, many values have been manipulated to improve the models. While this operation may have aided, it still involves altering the data on which the model relies.


THANKS FOR YOUR ATTENTION

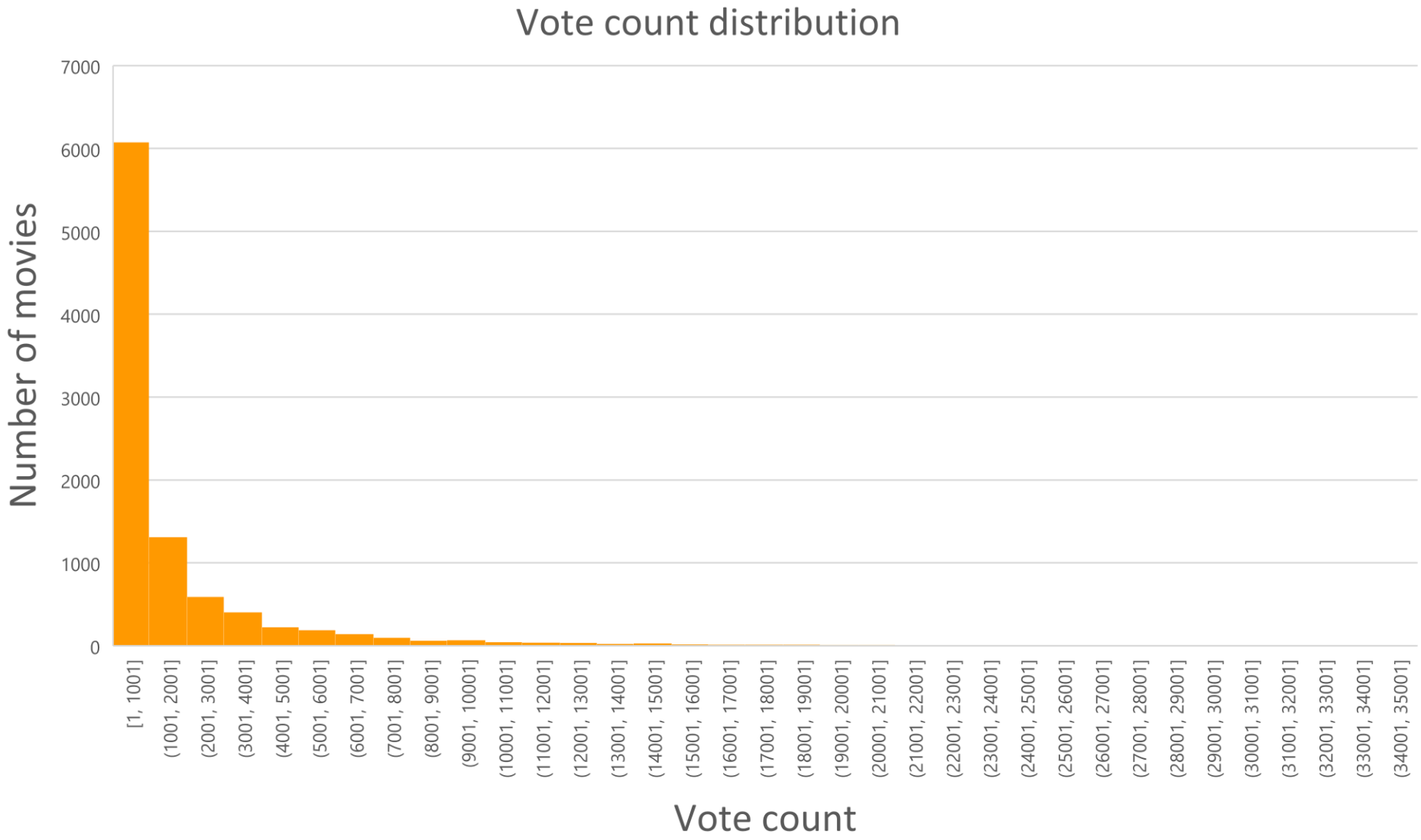

ANNEX

Appendix 1.1: Vote Average distribution



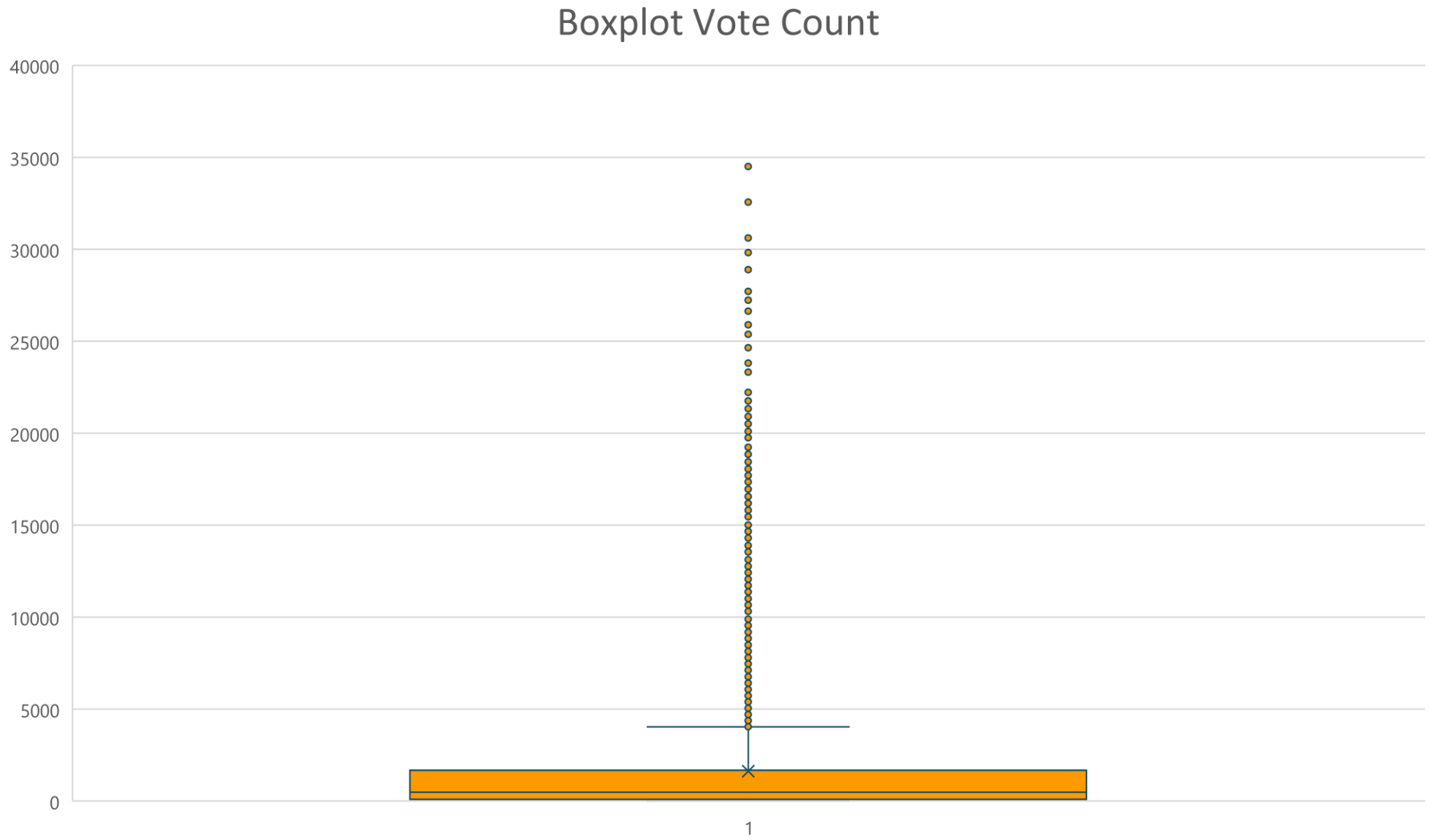
Bin: 0,2
Slide: 13

Appendix 1.2: Vote Count distribution



Bin: 1000
Slide: 13

Appendix 1.3: Vote Count Boxplot

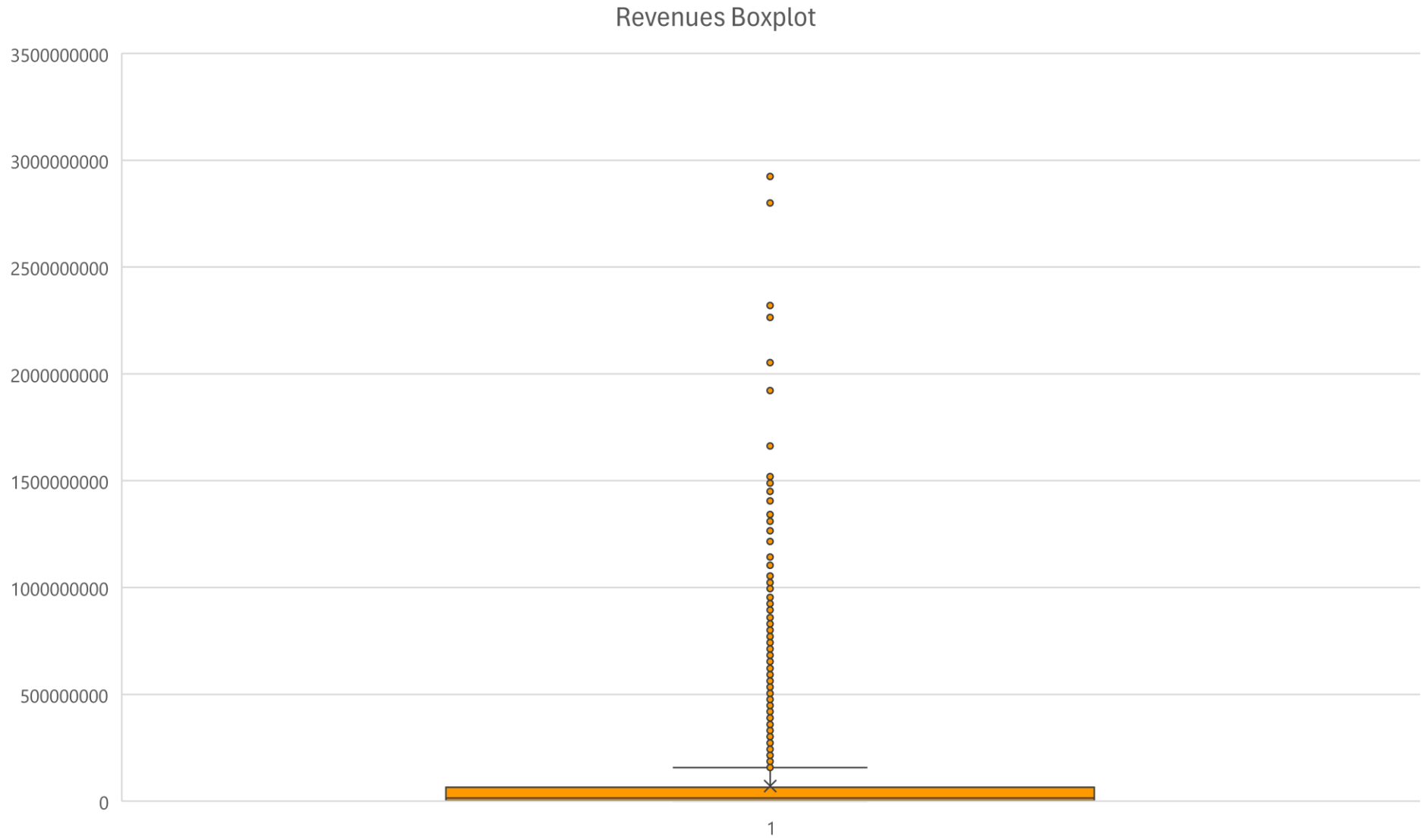


Slide: 13

Appendix 2.1: Vote count distribution

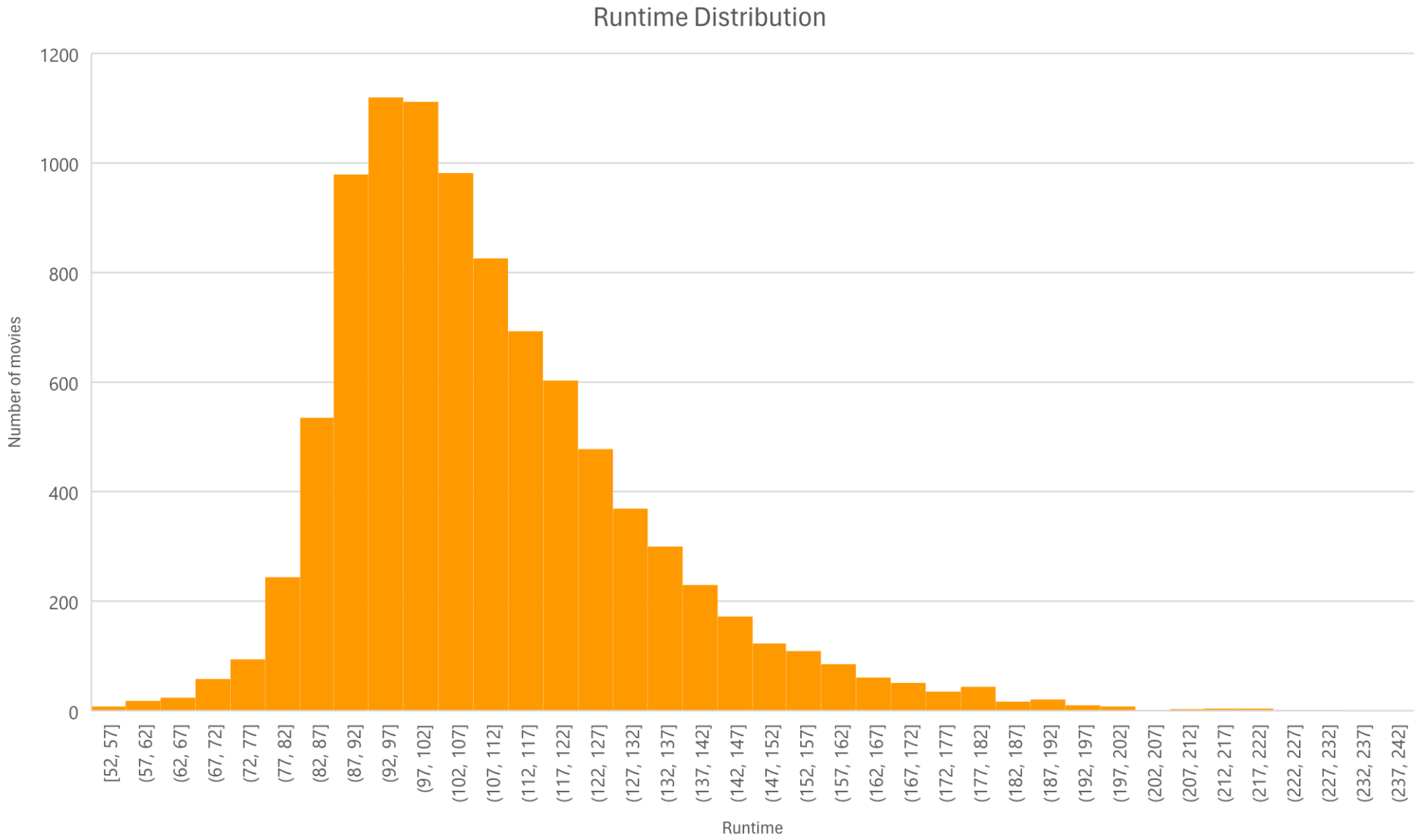
Bin = 260 M
Slide: 14

Appendix 2.2: Vote count distribution



Slide: 14

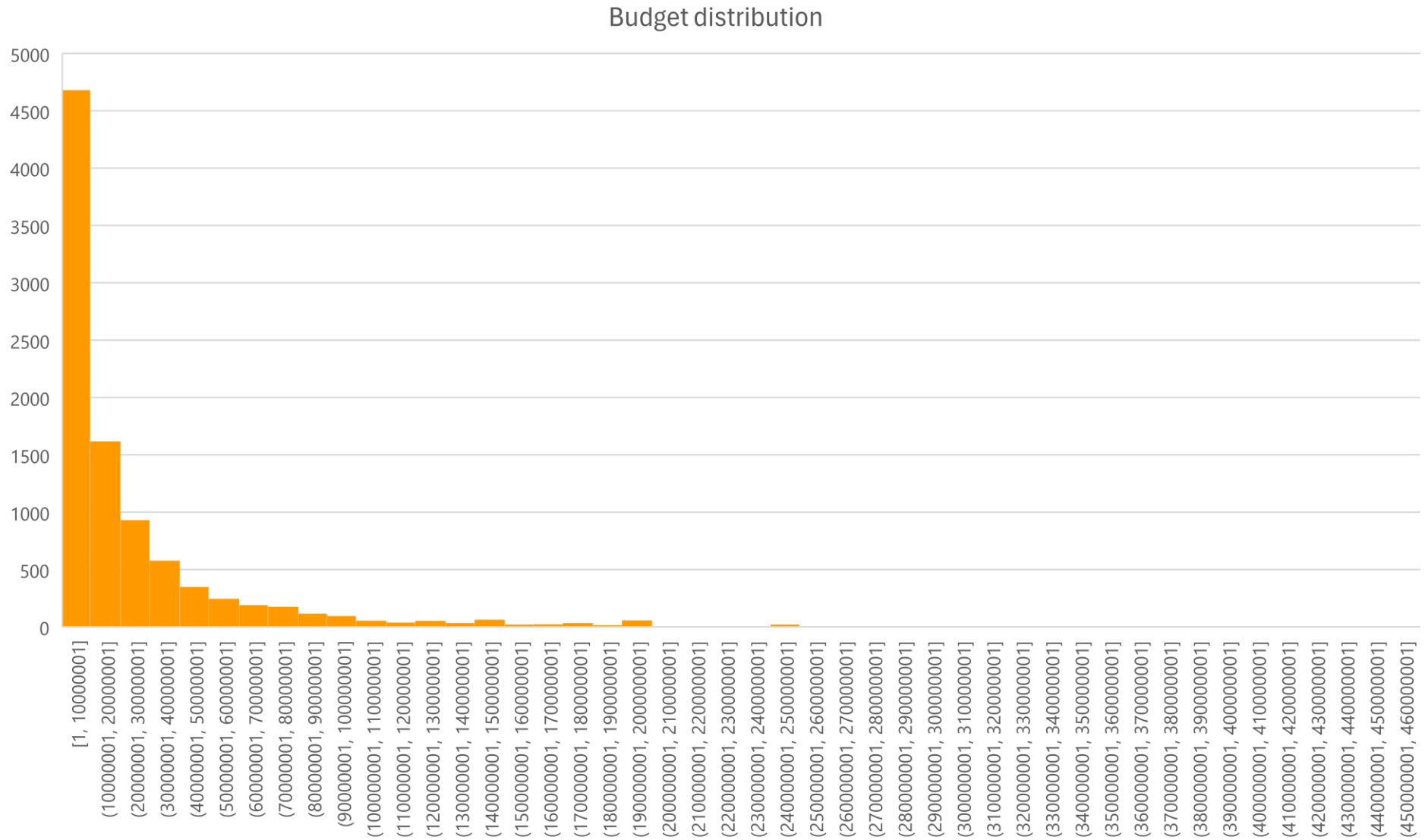
Appendix 2.3: Runtime distribution



Bin = 5 min

Slide: 14

Appendix 3.1: Budget distribution

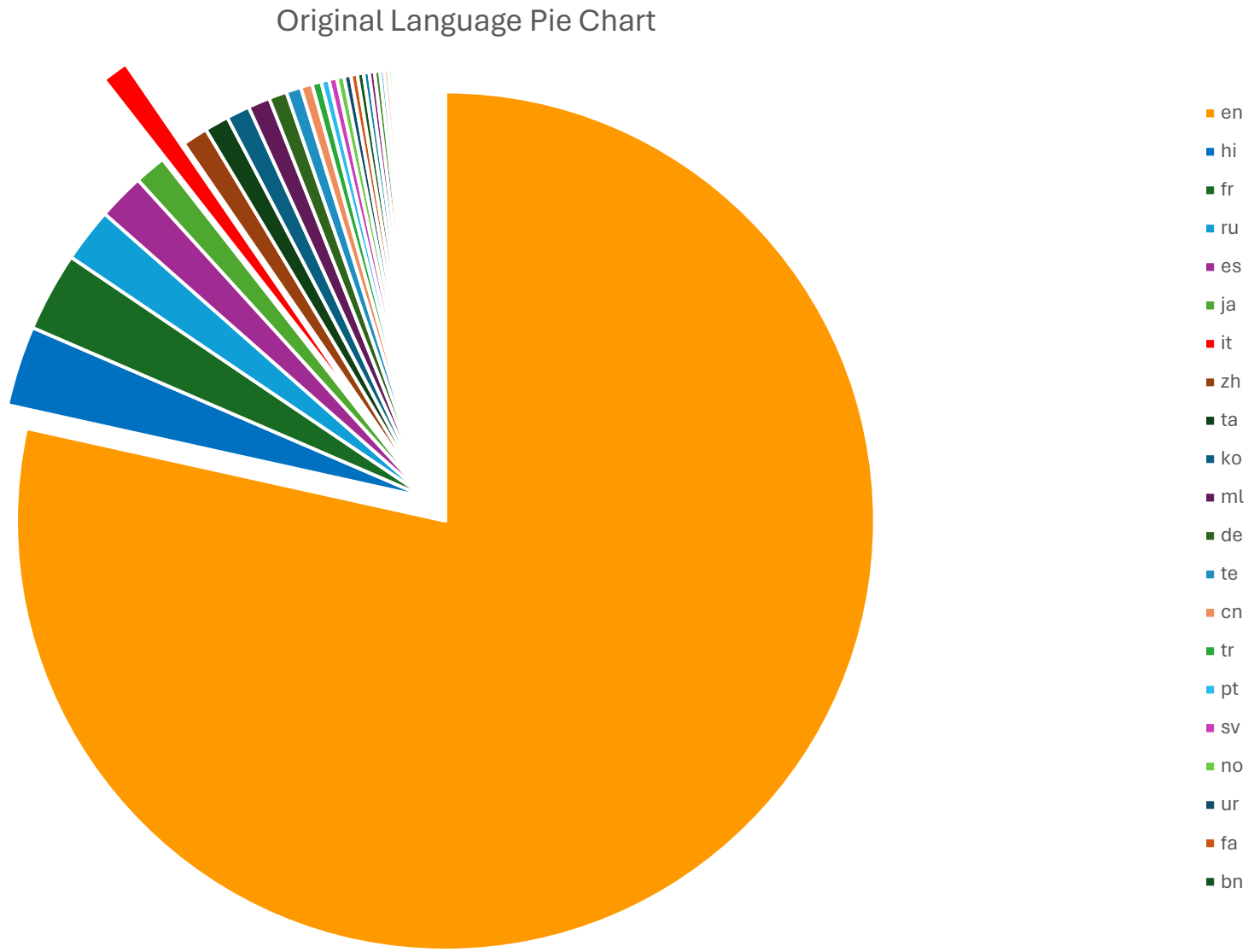


Bin = 10 M
Slide: 15

Appendix 3.2: Budget Boxplot

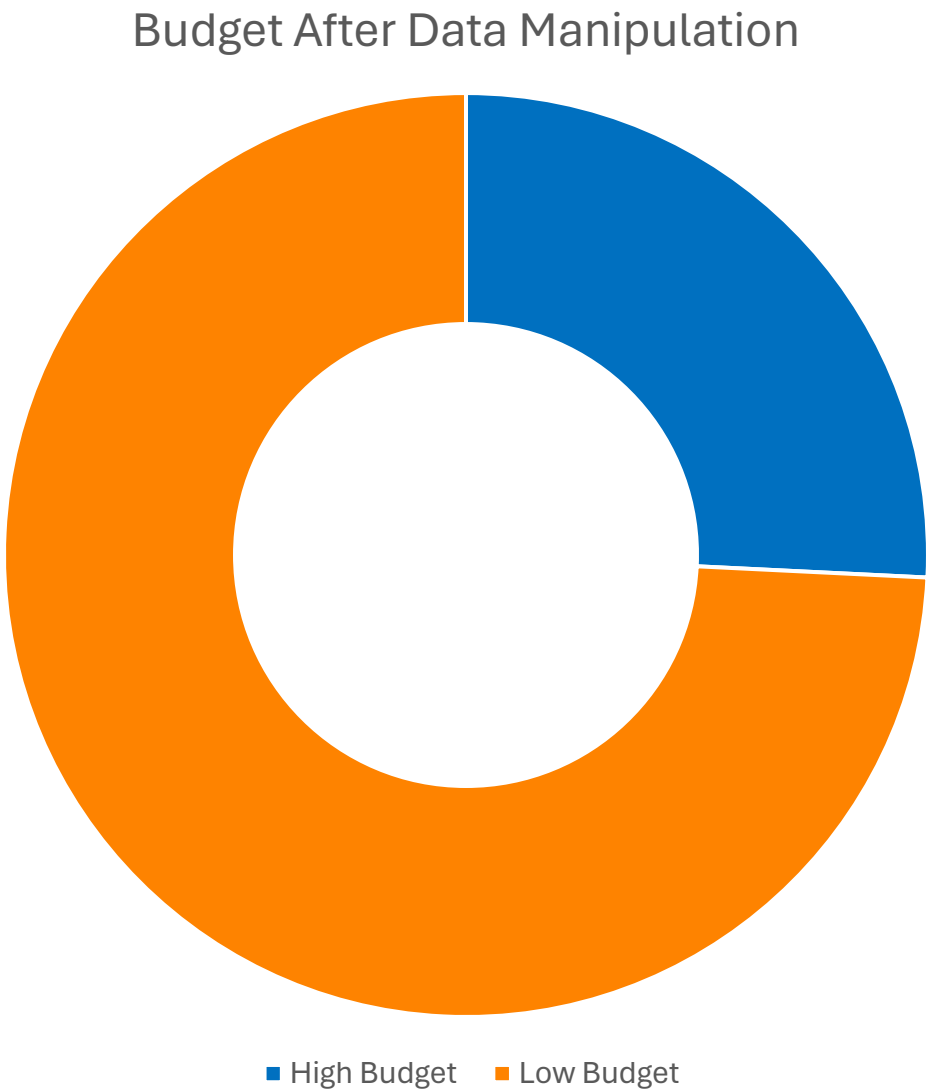
Slide: 15

Appendix 4.1: Original Language Pie Chart



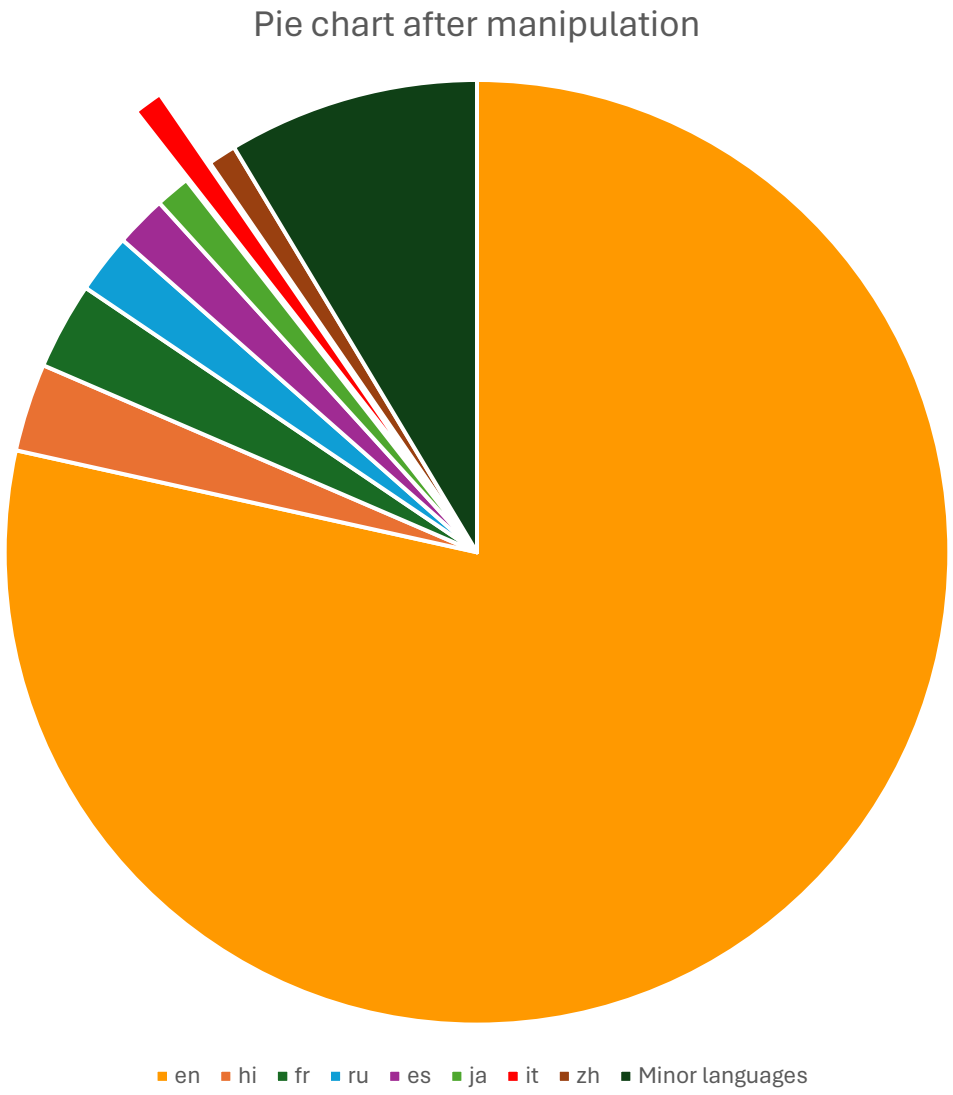
Slide: 16

Appendix 5.1: Budget After Manipulation



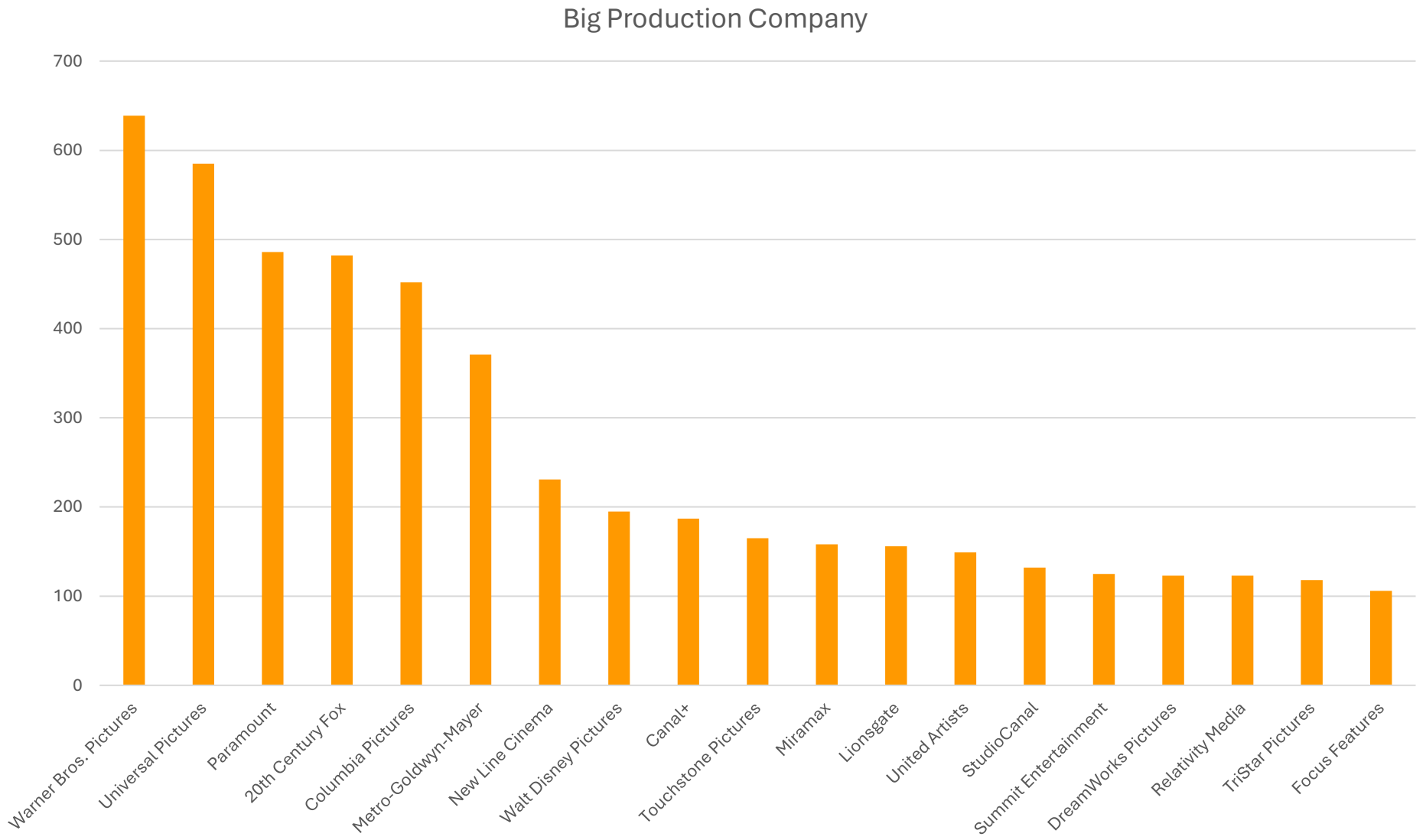
Slide: 22

Appendix 5.2: Original Language Pie Chart After Manipulation



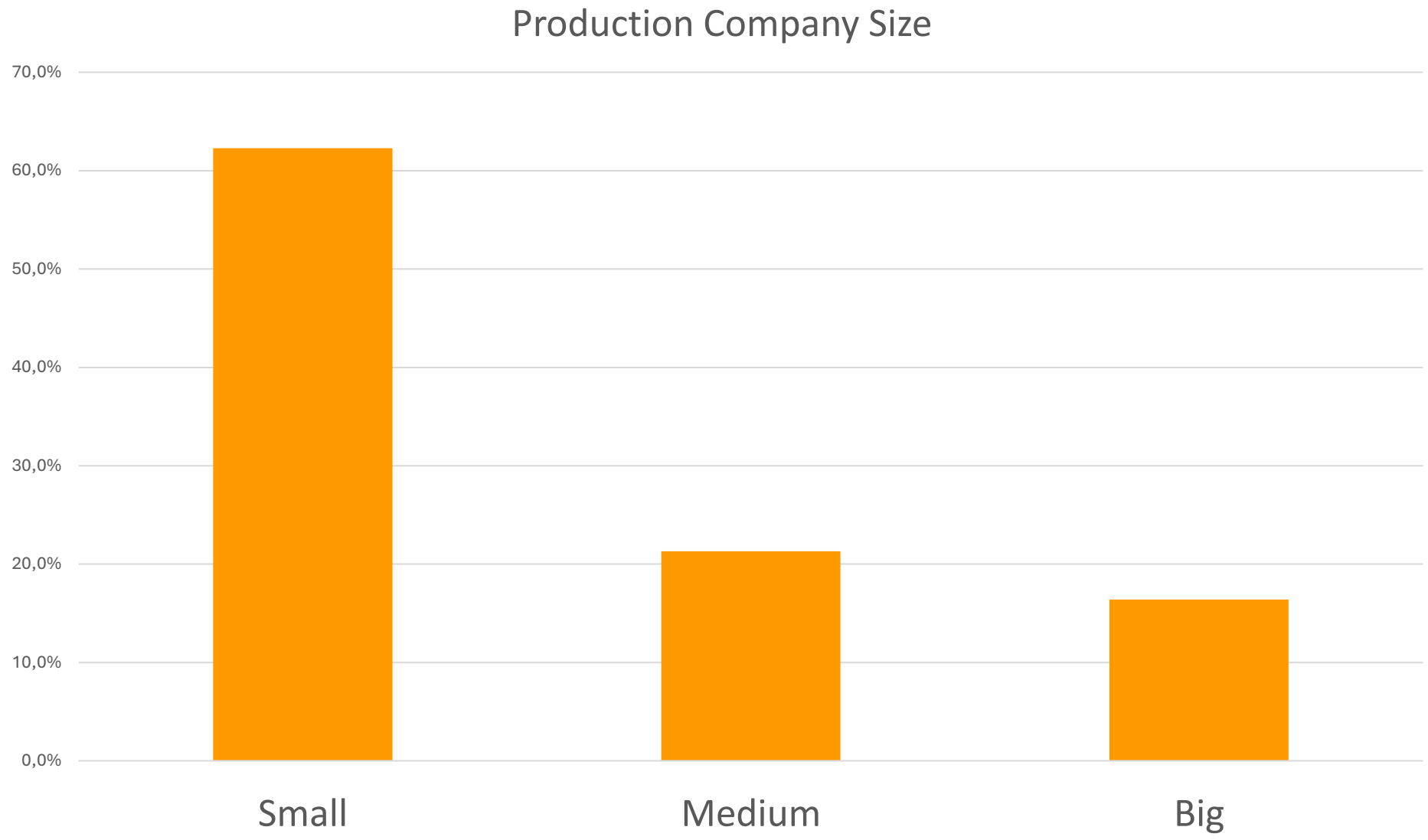
Slide: 22

Appendix 6.1: Production Company



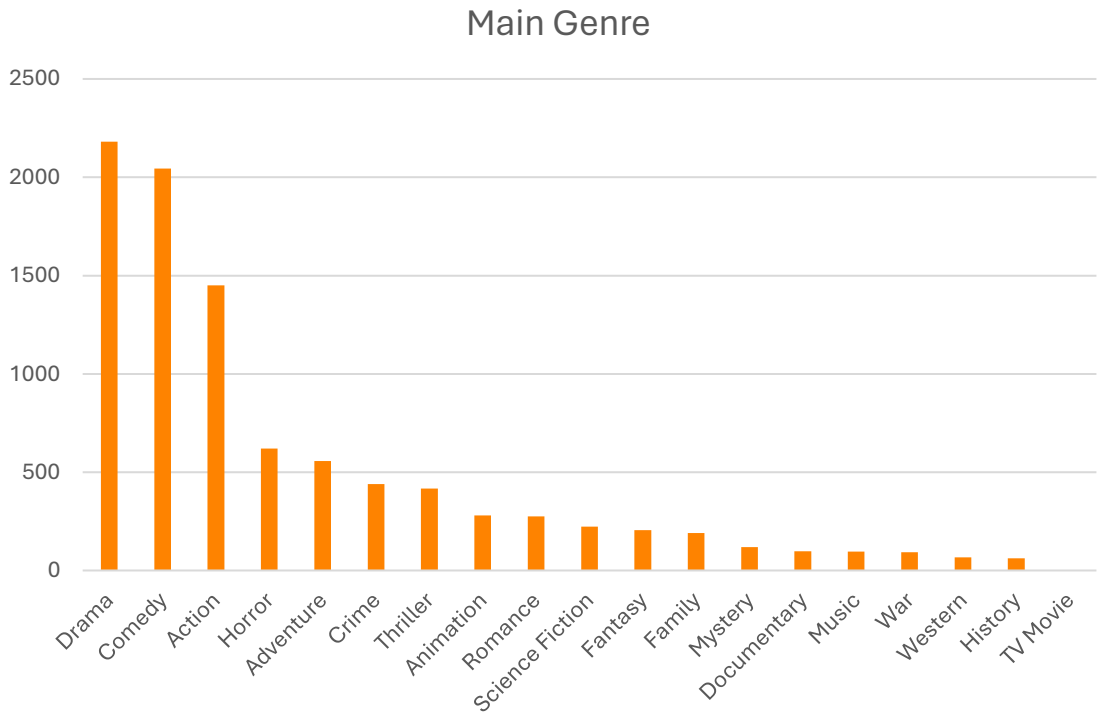
Slide: 23

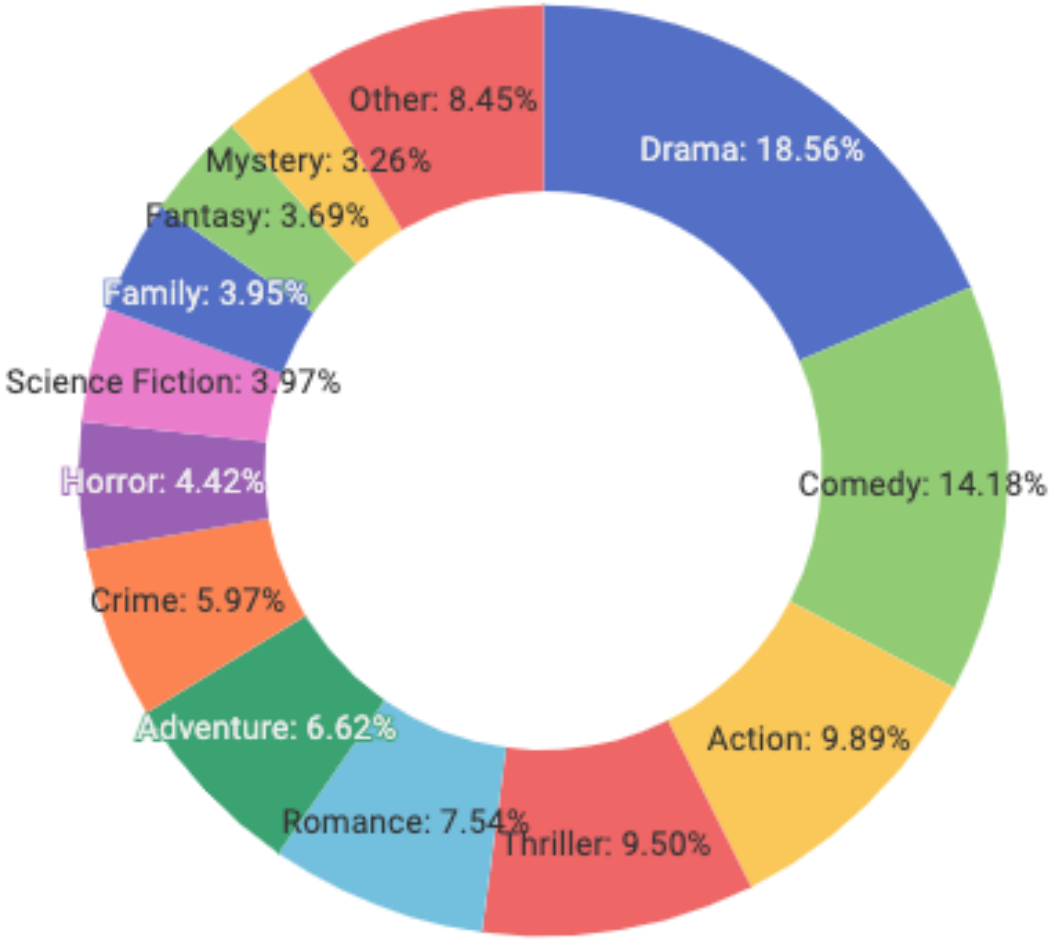
Appendix 7.1: Production Company



Slide: 24

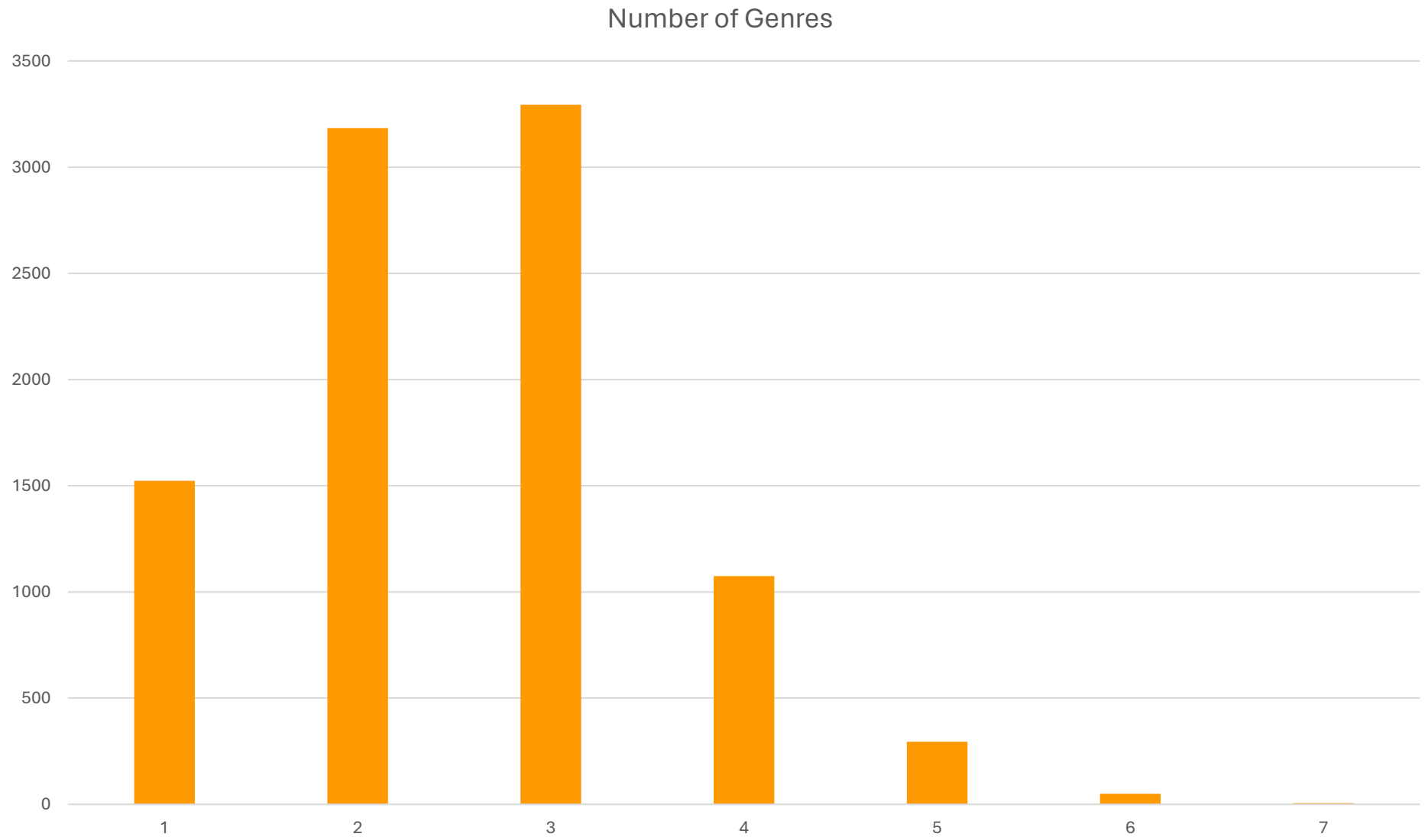
Appendix 8.1: Main Genre





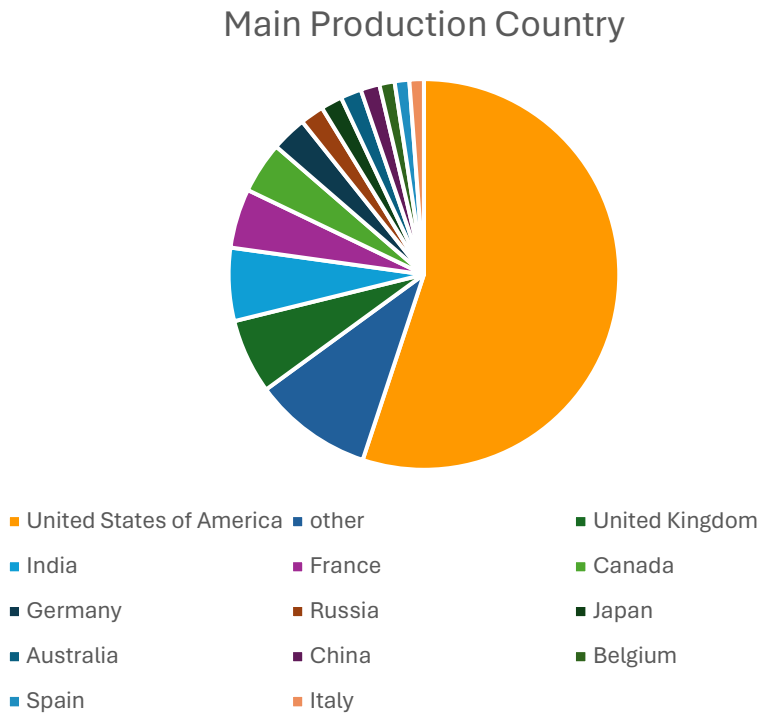
Slide: 25

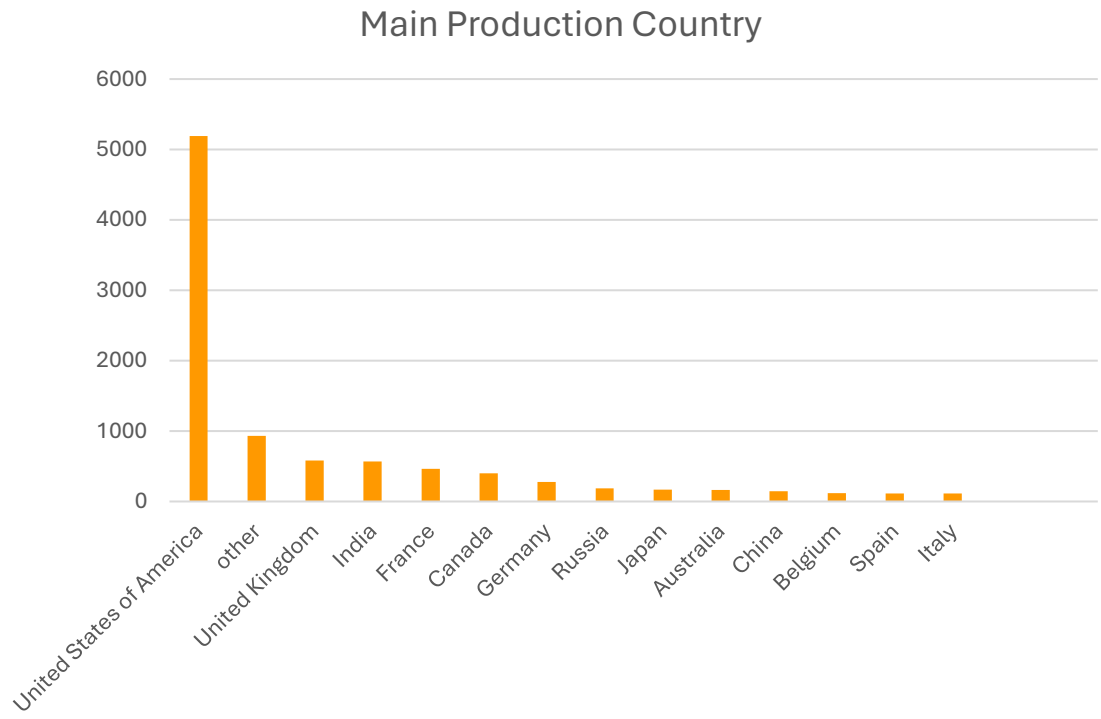
Appendix 8.2: Number of Genre



Slide: 25

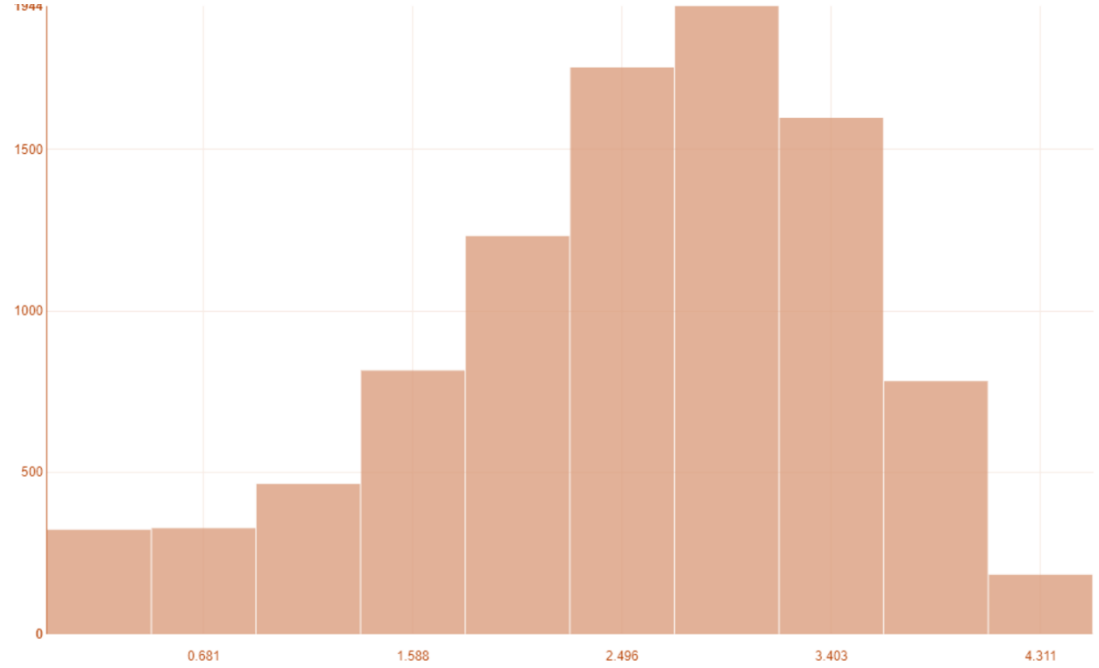
Appendix 8.3: Main Production Country



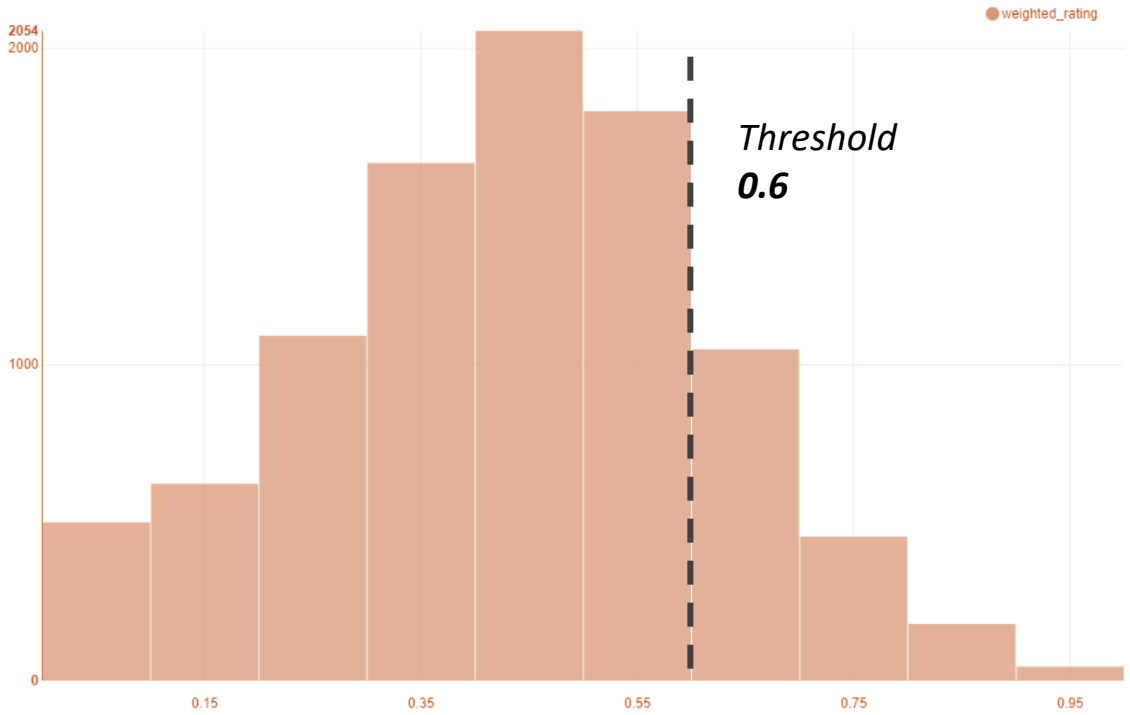


Slide: 25

Appendix 8.3: Vote count logged and Weighted rating normalized



Vote Count Logged Distribution



Weighted Rating normalized Distribution

Slide: 17