university of groningen

faculty of science and engineering

mathematics and applied mathematics

Analysing key factors affecting longevity of film’s screening in cinemas

Bachelor’s Project Mathematics

February 2024

Student: M. Stulenkova

Student number: s4325478

First supervisor: Prof. Dr. M.A. Grzegorczyk

Second assessor: Dr. Wim Krijnen

Abstract

The data analysis field nowadays is becoming commonly integrated in the film industry to gain insight on the success of their projects. This bachelor project applies survival analysis to the data of 905 wide released movies produced and screened in the US from 2013 to 2018. The aim is to explore the effects of production level and release stage film's characteristics on its success, where we define success as the amount of screening days. The methodology consisted of the Cox regression model with mixed effects, where we proposed the shared frailty for film distributors. In addition, we include the model validation and comparison with the standard Cox proportional hazards model. The results suggest that variables such as a film's rating, opening gross and opening theatres amount compared to maximum theatres affect the films survival. The effect of the running time is partly supported, while genres were not significant for the model.

Acknowledgement

Dedicated to my mother, for her immeasurable love and passion towards mathematics,
to my father for his enthusiasm for films,
to my beloved siblings whose presence in my life has made me a better person,
to Min Dat for believing in me and his patience.

I would like to express appreciation to Professor Marco Grzegorczyk for his guidance throughout the project and support with developing the ideas.

Contents

1 Introduction

The American movie industry nowadays is the largest film industry in the world. Commercially, it is the most successful: the total gross (domestic USA box office) for January 2024 counts more than 495 million dollars [IMDB, 2024].
Since the film industry has a high number of staff, actors, audience, the whole working process was significantly influenced by the COVID pandemic and its restrictions: e.g. the cinemas were mostly closed as it was impossible to follow the restrictions on social gatherings in indoor spaces.

This had consequences, that is, popularity of the streaming platforms, changes in film trends and the related post-COVID recovery. Streaming platforms such as Netflix, rely on data analysis, which allows for analysing and identifying trends in the film industry as well as viewer preferences in different movie characteristics.

Regarding post-pandemic period, the movie industry had experienced a decrease in box office gross, which resulted in mass reduction of staff. The aftermath of the pandemic resulted in one of the biggest writers strike interrupting film production for 5 months [Los Angeles Times, 2023]. The popularity of streaming platforms during the COVID-19 pandemic also raised a question of whether theatres/cinemas have become unnecessary even when the restrictions are lifted.

Currently, the trends show a decreasing amount of movies screened in cinemas. The reduction of the number of release movies is also observed in our preliminary analysis: in our dataset we calculated, that before the COVID pandemic the audience had a choice of 37 movies on average per day, while upon checking a sample of biggest USA theatres, we found that the ongoing releases allows for the viewer to choose 24 movies to see a day, and this is also confirmed in some of film journals, e.g. [NoFilmSchool, 2022]. Therefore, our research aims to analyse the factors which affect the length of the theatrical window of the films released in a period of 2013-2018[1], i.e. right before the COVID pandemic. We believe, that the results of this study are significant for performing further comparisons of the trends with the data sets from mid-COVID and post-COVID periods.

Before we state our research question, it would be beneficial to present the overview of the main literature, where the authors implemented survival analysis on the film industry data. The purpose is to search for the relevance of our chosen topic, as well as to aim towards the comprehensive research with the methodology and variables collected. Therefore below we provide the brief literature review of the researches connected to our work.

[Kim, 2021] applied survival analysis, namely Kaplan-Meier curves and Cox regression model, to examine factors that affect the number of screening days a movie stays in theatres, as the measure of the success of the movie. The data is from the Korean film database of 746 movies released during 2010-2019. The results of this study state that the domestic identity of the producer, positive, negative or neutral reviews have the significant effect on

[1] 2019 is not included, since the closing dates in theatres of the movies released in 2019 is affected with COVID-19 theatre closures

the movie's survival. The effects of supplementary factors, for the rating and genre, are partially supported (the effect on the survival time differs slightly for different genres and ratings).

[Legoux et al., 2015] also analysed the effect of critical reviews on the decision to keep or withdraw a movie from the theatre, taking the screening weeks as the survival time. The methodology is the implementation of basic discrete-time proportional odds (DTPO) survival model with random effects that were used for the theatres (118 theatres), to account for the heterogeneity. The variables tested are: Film gross and Screen numbers in prior week, Critical reviews, Rating, Genre, Distributors, Runtime, Stars, Production budget, Weekly consumer ratings, competitions and advertising. Their data consists of 788 films that were screened in Quebec, Canada between 2002 and 2011. The main source for this database was CINEAC, secondary are IMDb and Rotten Tomatoes. The results showed that covariates, which were similar to the ones considered in this research, such as distributor and runtime were not significant and did not affect the risk of the movie being withdrawn; the genre and rating were observed to have affect on the movie survival.

[Chisholm and Norman, 2006] use the survivor function to analyse the weekly film programming choices of three theatres in the Boston metropolitan market during the period of June 30 2000 - June 28 2001. The methodology involved the Kaplan-Meier product-limit estimator, Cox proportional hazard model and the accelerated failure-time (AFT) model. The data was obtained through Synergy Retail, IMDb and Hollywood Reporter. This paper considered variables such as: film's rank (based on gross relative to other films screened at the same week), film's first weekend revenue, the percentage of theatre gross, the number of opening screens during the first week of release, the "star power" of the cast and theatre owners. The results suggested that the covariate of the number opening screens do not affect the film's survival at the theatre level, while the first weekend revenue strongly affect the probability of stay.

We point out that there is a sufficient amount of researches on film's success, since the data analysis field in the film industry is growing. For instance, while performing the literature research we have observed other statistical methodologies such as linear regression models and machine learning techniques applied to explore the factors contributing to a film's success. However, the survival analysis usage is rather limited in the amount of researches being conducted on the movie data, where the screening in theatres is defined as the survival time.

This paper aims to answer the following research question: **"To what extent do movie production characteristics and release factors affect the success of movies?"** by using the survival analysis. Movie production characteristics refer to the predetermined information about the movie: distributor, genre, rating, running time. First-day and post-release factors then refer to the data such as the first day opening gross and the proportion of open theatres compared to maximum amount of theatres which screened these movies. We measure success of the movie in the amount of screening days in theatres, defining survival time as the amount of screening days. Moreover, this paper attempts to improve the research methodology of implementation of the survival analysis, specifically by using the extended Cox model and the inclusion of frailty for the Distributor variable.

Based on the definitions of ratings, we formulate the hypothesis as following: **Hypothesis 1: The rating of a movie has an effect on its amount of screening days.** We further hypothesise that the PG & G ranking movies will have a higher probability of stay in theatres as time passes and as we explain in the Data section, the main reasoning behind is that the movies which have no age restriction attract bigger audience, thus, are more successful.

Regarding the genres of the movies, we have **Hypothesis 2: Genre is an indicative characteristic of the movie in terms of its success.**

For the running time, it is widely discussed that there is a trend for an increase of overall length of movies [Rosser, 2019]. Hence we investigate **Hypothesis 3: The runtime of a movies affects its longevity of stay in theatres.** Here our prediction is that movies of a longer runtime will have a higher probability of stay in theatres compared to those of shorter or average runtime.

Since we investigate the success of the movies, the amount of money that the movie makes during its first day (the opening gross) is an indicative characteristic of its future success. Hence our **Hypothesis 4: The movie's opening gross amount has an effect on its longevity of stay in theatres**. Our prediction is that movies with higher opening gross will have a higher probability of stay in theatres as time passes.

For the amount of opening theatres, we aim to explore whether the comparison with the amount of maximum theatres, that is, the growth of the wider release (movie screened in more theatres as time passes) contributes to the success of the movies in terms of the screening days. Therefore, we explore **Hypothesis 5: An increase of movie's screening theatres affects its longevity of stay in theatres.**

Regarding the distributors, which are the studios producing the movies, we assume that there is a within-cluster correlation for movies belonging to the same distributor. We further explore whether the choice of the distributors as random effects is justified.

The structure of the work is divided into data section, the inference on the Extended Cox model, the model assessment and conclusion&discussion. We finish the main paper body with the model comparison to the Cox proportional hazards model and perform model assessment and validation by backward elimination, cross-validation and assessment of normality assumption.

2 Data

Within this section we present the data sources used in the paper and provide the details about the data collection, cleaning and formatting processes, together with the background and motivation for the categorization of variables. The collected data sample required detailed and careful preparation for the analysis. The data set contains the films produced and released domestically in the US, which is a commercial industry, thus we point out that a lack of transparency and variety of open source information is present to some degree. But we believe that our data sources are reliable given that they are widely used in other researches.

2.1 Data manipulation

It is desirable to choose the movies that are accessible to a wide population, hence eliminating limited and "exclusive" releases [OpusData, 2023]. Since we focus on a commercial side of the film industry, investigating the success of the movies, we exclude "experimental" projects in the limited release which are less motivated by commercial success. Moreover, the variables in the limited release data can be difficult to categorize: i.e. the characteristics such as runtime, genres and opening gross, since they differ from the "standard" (wide release) films. That way we believe that a more comprehensive inference can be extracted with the wide release movies, i.e. screened in larger than 600 theatres.

2.1.1 Data collection

The three main sources of data are IMDb Box Office Mojo [IMDB, 2024] and Nash Information Services [Nash, 2023] as the prior data sources, The Numbers [Numbers, 2023] as a secondary source.

Since the main methodology of this research is application of survival analysis, we first collected data from IMDb Box Office Mojo[2], which contains the number of screening days (days of film's stay in theatres), defined as survival time of this research. In addition to the dates of release and closing in theatres, a few other variables were collected from IMDb as summarized in Table (1).

In order to increase the amount of collected variables, we contacted several other movie industry data organizations. Our request was addressed by Nash Information Services, who provided access to a dataset intended for academic usage[3]. Upon merging the extract with IMDb data set, we found out that approximately 1/3 of the number of movies were absent from Nash. Unfortunately the data sent by Nash was the subset of the full data, therefore some of the variables provided there (creative type of the movies, production method, etc) could not be used for our type of the research, since that information was not found to be available in other sources, to manually collect it for the unmerged movies.

[2] Information courtesy of IMDb (https://www.imdb.com). Used with permission.

[3] We hereby acknowledge the usage of the intellectual property of Nash Informational Services in this publication.

After we merged the data from IMDb and Nash, the unmerged subset of data was collected manually from IMDb, for the variables present there, using different directory on the website and The Numbers was used for the reliability of the data collected. For some of the variables missing, we also collected the information using the same databases, but different directory on the websites. To ensure the reliability of the collection, the sample of movies was rechecked in The Numbers, so it served as the supporting database.

To summarize, the total amount of movies from IMDb Box Office Mojo before data cleaning and further data inspection consists of 987 films. The total number of movies accessed from Nash Information Services upon request - 1900 for years 2006-2018, as the production years of the films[4]. Thus, we examined that from the production years 2006-2009 there is no films that matched with ones in the IMDb dataset and the production years of films which were included in this research started from 2010 onwards: starting with 1 match in 2010, 24 matches were found in 2011 etc. Below is the table summarizing each of the variable collected data sources.

Variables	**Source**
Distributors; Opening Gross; Amount of screening day; Open Theatres; Maximum theatres	IMDb's Box Office Mojo
Genres; Rating; Running time	Nash Information Services; IMDb's Box Office Mojo and The Numbers

Table 1: Data sources of variables

2.1.2 Data cleaning and formatting

In the data provided by Nash, the rows (corresponding to each film) were initially contained in one cell each. Once this was resolved, the film names were split if the commas or other special characters were present in the film titles. Thus the formatting issues were present and had to be resolved.

Regarding the formatting for IMDb set, the Kutools extension was used, together with the built-in Excel functions for solving the issues. For 11 movies in total, the closing day was not available, hence for those movies we calculated the screening days by tracking the daily gross information on IMDb (where the closing date of the movie in theatres should have corresponded to the last information about the gross). We converted the information into the date format, enabling calculation of the amount of screening days. When the function for calculating the amount of screening days was used, it was observed that the count between the dates did not include the release day itself, so +1 day was added to this function. The

[4] Production years differ from the release years, usually production took place a year before the release dates. Production year corresponds to the movie-making period.

numbers/currencies also were converted to the correct types.

After solving the formatting issues for both data sources, we merged the sets, resulting in merged 654 films. We also excluded the duplicates (in case the movie was released and re-released in period 2013-2018) and formatting the names of the movies to match (if the movie name contained symbols it was wrongly perceived by Excel). The next step was to manually collect the remaining data for unmerged movies, ensuring that the collection is accurate with our previous corrections. We also note, that here by "manual" collection we mean the collection of variables not from the convenient table of movies of data on the IMDb website, but mostly for every movie separately.

Since we need feasible data to work with, we continue with careful examination of the data set. Some characteristics of the movies affected their inclusion in this thesis. Importantly, in the IMDb dataset it was observed that some of the movies present are re-released. After examining the distributors, it was noticed that Fathom Events is the distributor of re-released movies, ones produced by other studios, and "event" movies, e.g. theatrical plays shown in theatres, so the movies distributed by this studio are not relevant for the analysis. As described in the Introduction, the point of our research is to analyse the films originally released in 2013-2018. Thus, we explain the exclusion of re-released movies by the variables, such as opening gross and opening theatres, be different from its original release, which would bias the results. Similarly, we excluded all movies and distributors that belong to the re-released and experimental category.

After getting rid of re-releases, Fathom events movies and some experimental movies, we have a final sample of 905 movies, with the complete information for each variable.

2.2 Variable Description

For this study, we think that the reader has to be familiar with some topics in the film industry. Concepts such as the relationship between distributors and theatres, basics of the genre theory and terminology for other variables in the film industry are referenced for better understanding and correct interpretation of the variables.

The process of delivering a movie to the audience is divided into three main stages: production, distribution and exhibition. We consider variables for each stage: genres, ratings, distributor and running time for the production stage, opening theatres for distribution, and opening gross and the amount of screening days for exhibition stage. Upon preliminary analysis of data it was decided to categorize some variables, and we provided the corresponding explanation to the categorization according to the film literature and magazines.

2.2.1 Distributors

Distributors are the film-making studios, which set the film's release dates and negotiate with theatres the theatrical windows [Pokorny et al., 2019]. The initial theatrical window is agreed upon by both parties. In case a movie performs well commercially, the theatrical window can be extended.

In the film distributor industry in the US, some studios are leading: Walt Disney, Warner Brothers, Universal, Sony Pictures and Paramount Pictures. For this research, we did not group the Distributors into subcategories, however, it is possible to group them according to the market shares. Hence, our analysis involved all 57 distributors as separate categories [IMDB, 2024]. We assume that films which have the same distributor at the production stage have common characteristics unobserved in our data such as financing, staff, shared facilities etc.

2.2.2 Genres

When we collected the data, we decided to take the predominant genre of the movies, since some movies had several genres in one data sources, and only one genre in the other. Thus, we searched the definition of the genre, as a movie characteristic, and we note that the classification of movies based on genre is complicated. According to [Altman, 1999]: "A genre is not an average descriptive term, but a complex concept with multiple meanings, which can be identified as following: as a formula, a formal network of films, a category and its perception by the audience."

Due to the different definitions and interpretations, genre theory is a topic which complicates the process of variable categorization in this study. For example, we can group Adventure and Western genres because they have similar characteristics of storytelling and dynamics of the plot, but grouping Drama and Comedy is not appropriate, since they are different by characteristics.

For the Documentaries group, the Biographies and Documentaries can be grouped in a single category due to the similar definition of the real-life oriented movies. Regarding the Comedy genre, to this category we combine: Black comedies, Romantic comedies etc., they are all grouped into a main genre since they all have a similar entertainment characteristic. Musical and Concert/Performance genres are grouped into the same genre due to similarities in music theme. Finally, the Thriller genre has similarities with Crime and Suspense genres in having criminal topics and a suspense atmosphere.

To summarize, we categorized the genres into eight groups (in the bracelets we indicate amount of movies and initial genres): Action (161), Adventure (adventure + western) (143), Documentaries (documentaries + biography) (48), Comedy (black comedy + comedy + romantic comedy) (192), Musical (musical + concert/performance) (12), Thriller (crime + thriller + thriller/suspense) (92), Drama (194) and Horror (63).

2.2.3 Rating

A rating given to a movie, according to [MPA, 2020], refers to the audience which is allowed to access the screening in cinemas. The age limit on a film can affect the size of a potential audience, hence we suggest it affects the success of a movie in cinemas. Ratings for movies are assigned by Motion Picture Association of America (MPAA), which act as a guideline for the parents to consider whether a movie is suitable for viewing by their children, including the presence of mature themes (language, depiction of violence, nudity, sensuality, depictions of sexual activity, and drug use). Below is given the description for each restriction.

- G - General Audiences. All Aged Admitted
  Contains nothing of the mature themes, in the view of a Rating Board, would be inappropriate for children of any age.

- PG - Parental Guidance Suggested. Some Material May Not Be Suitable For Children
  The PG rating indicates, in the view of the Rating Board, a PG rated movie should be investigated by parents before they let their younger children attend the screening. Audiences of all ages are still allowed to attend the screenings, since no objective mature themes are shown.

- PG-13 - Parents Strongly Cautioned. Some Material May Be Inappropriate For Children Under 13.
  A PG-13 motion picture may go beyond the PG rating in theme, but does not reach the fully restricted category. There may be depictions of violence in a PG-13 movie, but generally not realistic. Children under age 13 are not allowed to enter unaccompanied by parents or guardians.

- R - Restricted. Children Under 17 Require Accompanying Parent or Adult Guardian.
  An R rated movie may include adult themes, so that parents are counseled to take this rating very seriously. Children under 17 are not allowed to attend R-rated motion pictures unaccompanied by a parent or adult guardian.

We note that the restrictions and their definitions may be different in Europe. Moreover, upon examining [MPA, 2020] we noticed that the ratings of movies can be changed if there are two versions. As MPAA states: the original release rating, if it was in the restricted categories, after the end of screening period, can be changed, that is, the adult content from the movie can be edited/removed. The movie can then be re-released with different age restriction. The current dataset involves only the ratings assigned for the original release.

According to the above definitions, we group PG&G ratings into one category as the parental advisory recommended but not required. We categorized ratings as: R (367), PG-13 (390), G&PG (148).

2.2.4 Opening Gross

For a more comprehensive study of the opening gross, we note, for further studies, that the year, economical factors and the production budget could have improved the insight for this variable. We believe that the opening gross is an important indicator of the film longevity in theatres, since the gross itself is used as the measure of success of the movie.

The range for the Opening Gross in this data is significant: from $10,723 to $257,698,183. Considering the above, in order to get more comprehensive insight for the influence of the Opening gross values on the film's duration in cinemas, we decided to categorize the opening gross into the equal-sized groups (divided into quantiles). That way, we believe the theatres would have more insightful information in order to predict the success of the movie based on which category (quantile) it belongs.

The financial success of a movie during its opening days can affect the decision of a theatre to extend the film's screening. The responsibility for a film's stay in release belongs to the distributor and the theatre. Without the production budget variable, our categorization is mostly informative for the theatres, than for the distributors, who produced the movie. As amount of screening days increase, the rights for closing a movie's screening transfers to the theatres, i.e. the theatres have the decision making power to withdraw or continue film screenings depending on their profitability [Pokorny et al., 2019]. The dataset was divided into quantiles as following:

Category	Movies	Corresponding Opening Gross
Very Low ($<$ 20%)	181	\$10,723 - \$1,729,002
Low (20-40%)	181	\$1,729,002 - \$7,596,687
Below Average (40-60%)	181	\$7,596,687 - \$15,271,843
Above Average (60-80%)	181	\$15,271,843 - \$29,872,748
High ($>$ 80%)	181	\$29,872,748 - \$257,698,183

Table 2: Movie Distribution by Opening Gross Categories

2.2.5 Running time

The running time of a movie, based on the data collected, ranges from 75 min to 201 min. Below we divided the runtime into 10 min intervals to study the relationship between the average screening days for the corresponding intervals, Figure (1)[5].

We do observe the general tendency of the film's length in relation to the greater average amount of screening days, being longer, but the categorization into bigger runtime intervals is needed. The categorization for the running time could be done in various approaches. For the accurate investigation of our **Hypothesis 3: The runtime of a movies affects its longevity of stay in theatres**, we propose two approaches to either confirm or reject the hypothesis.

The first categorization we propose is based on the existing research, described in the Introduction [Legoux et al., 2015]. Given the skewed distribution of runtime (1), we divide the running time into bigger intervals, namely into 6 groups, based on the corresponding movie amount. We note that the shorter intervals follow from the most movie length being closer to the average runtime. We suggest, that the detailed examination could provide better

[5] visualization in Excel for preliminary analysis

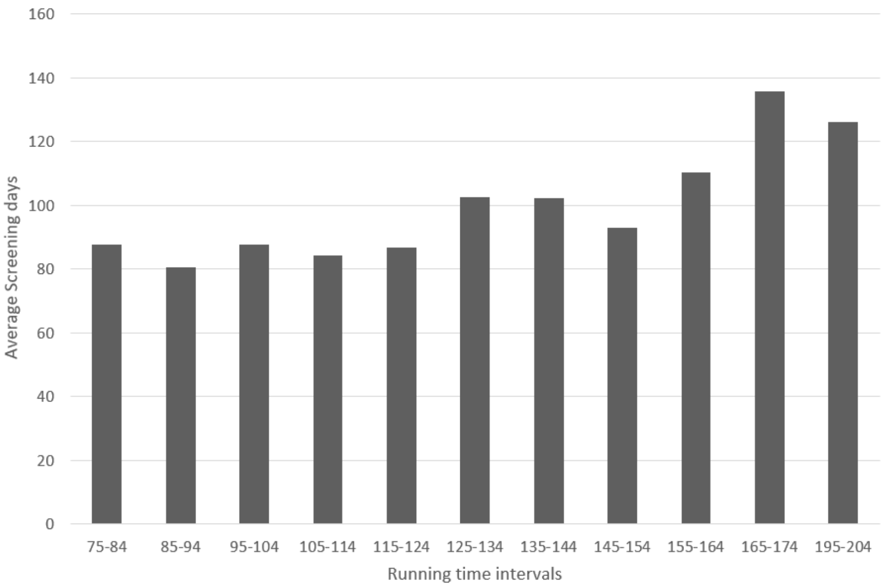

Figure 1: Running time preliminary analysis

model fit, since we could observe the corresponding tendency for the movie to be withdrawn from theatres, as the runtime of the movie increase/decreases (or observing no tendency).

Category	Movie Length	Amount of Movies
Group 1	75-94 min	159
Group 2	95-101 min	157
Group 3	102-108 min	146
Group 4	109-116 min	150
Group 5	117-126 min	143
Group 6	127-201 min	150

Table 3: Categorization 1 of Movies by Length

The second approach we propose is the division into 3 categories: Short length, Average Length and Long Length movies, Table (4). This categorization is appropriate for validation of **Hypothesis 3**, also addressing the skewness (by having better average screening days differences between categories) but there is a limitation on the better trend examination of the relationship between the increase of movie length [Rosser, 2019] and corresponding success.

Category	**Movie Length**	**Amount of Movies**	**Average Screening days**
Short Movies	75-99 min	267	82
Average-length Movies	100-124 min	475	87
Long Movies	125-201 min	163	103

Table 4: Categorization 2 of Movies by Length

2.2.6 Open Theatres

When releasing a movie, distributors arrange a certain amount of theatres which will have rights to display the movie. In this section we are interested in how the expansion of the amount of cinemas screening a film affects the longevity of its stay in theatres. Since we do not have detailed data on the progress of a movie's theatre presence, we take a percentage of opening theatres to maximum number of theatres. Similarly to the Opening gross description, the theatres observe the success of the movies based on the progress of its gross. Thus, if observing the success during some period, theatres can extend the theatrical window to increase the gross from a movie. Hence, given that half of movies in our dataset are screened in a maximum amount of theatres on its release day, we categorised them into groups with respective amount of movies suitable for analysis.

Open Theatres to Max. Theatres	**Movies**
$< 25\%$	142
25% to 90%	40
90% to 100%	294
100%	429

Table 5: Grouping of Movies by Percentage of Open Theatres to Maximum Theatres

2.2.7 Amount of Screening Days

The dependent variable in this paper is the total amount of screening days of a film in theaters, defined as the survival time. While some studies use total gross to measure success of movies, [Kim, 2021] suggested for the number of screening days as the accurate measure of success. [Legoux et al., 2015] also measures the success of the movies by its duration in theatres, but in screening weeks. We proceeded with the days count, since we believe it can better reflect the accuracy for the research. The count starts at the release date and finishes at the last day of a film in theatres, the closing day.

Below is the overview of the variables collected with the corresponding categorization.

Variable	**Categorization**	**Explanation**
Distributors	57 separate Distributors	We leave the Distributors as single separate categories
Opening Gross	5 Categories: Very Low ($10,723 - $1,729,002), Low ($1,729,002 - $7,596,687), Below Average ($7,596,687 - $15,271,843), Above Average ($15,271,843 - $29,872,748), High ($29,872,748 - $257,698,183)	Data is divided in equal sized parts based on the number of movies. The motivation is the lack of economical/production budget information and a huge range of the Opening Gross values. The grouping is somewhat subjective
Percentage of Open Theatres to Maximum Theatres	4 Categories: below 25% ; 25% to 90%; 90% to 100%; 100%	We take a percentage of opening theatres to maximum number of theatres
Amount of screening days	No need for categorization	Dependent variable; Survival time
Genres	8 Categories: Action, Adventure, Documentaries, Comedy, Musical, Thriller, Drama, Horror	The genres are grouped according to their basic common characteristics. The grouping is somewhat subjective
Rating	3 Categories: R, PG-13, G&PG	The covariates are grouped according to the parental advisory description provided by MPAA
Running time	**Categorization 1**: Group 1 (75-94 min), Group 2 (95-101 min), Group 3 (102-108 min), Group 4 (109-116 min), Group 5 (117-126 min), Group 6 (127-201 min) **Categorization 2**: 3 Groups: Short movies (75-99 min), Average-length movies (100-124 min), Long movies (125-201 min)	We propose two categorizations, in order to address the skewed distribution and to improve the inference on the Runtime influence. The grouping is somewhat subjective

Table 6: The Overview of Categorization

3 Statistical methods

This section covers the necessary mathematical preliminaries for the survival analysis, including the Cox proportional hazards model, then introducing the extension of the model, Cox mixed effects model with shared frailty, which includes the estimation procedure for the both, fixed and random effects. This section builds the theoretical framework for the analysis performed on the data gathered in the next section.

3.1 Survival analysis

3.1.1 Preliminaries

Survival analysis is the collection of statistical methods of time-to-event analyses, where the outcome (or dependent) variable is the time-to-event. By this, we mean that the time until the event occurs is measured and defined as the survival time. Referring to our research, we measure in days the time from the release date of the film (exposure of an event), until the end release date, thus "events" of interest are the amount of days the film exits the theatre (amount of screening days).

Survival analysis is used in the broad fields, i.e. medical researches, where the event of interest is defined as deaths of an individual and the survival time is the time until the deaths of an individual. In the survival analysis it is also possible to account for the cases when some individual survival time is unknown, referred to as the censored data. In our research, the screening days are measured for each movie, therefore we do not have the censored data, we have a complete data.
Next we introduce the basic terminology and notations in the survival analysis [Kleinbaum and Klein, 2005].

Definition 1.1. [Survival function]
The survival function given by

$$S(t) = P(T > t) \tag{1}$$

is the nonincreasing function, representing the probability that the individual's survival time, denoted as random variable T, exceeds any specific value of time t. The key characteristics are:

- $0 \leq S(t) \leq 1$
- at time $t = 0$, $S(0) = 1$
- at time $t = \infty$, $S(\infty) = 0$

Definition 1.2. [Hazard function]
The hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{2}$$

is defined as the instantaneous probability of the event of interest to occur in the time unit between t and $t + \Delta t$, given the survival up to time t.
Since the hazard function returns the probability per unit time, we refer to h(t) as conditional

failure rate, not as the conditional failure probability.

To summarize, we provide the relationship of S(t) and h(t):

$$\begin{aligned} h(t) &= -\frac{d}{dt}\log S(t) \\ S(t) &= \exp\left(-\int_0^t h(u)\,du\right) \end{aligned} \tag{3}$$

For more detailed preliminaries, the reader is advised to consult [Dobson and Barnett, 2018], if needed.

3.1.2 Kaplan-Meier

Plotting the survival function $S(t)$ over the time t, produces the survival curves, also known as the Kaplan-Meier curves. Given the actual data, the survival curves have the shape of the step functions, where occurrences of the events are represented as "steps". The Kaplan-Meier curves allow to compare the survival probabilities at each time unit and the failure rates by analyzing the slopes for different plotted groups.

3.1.3 Cox proportional Hazard model

Definition 1.3. [Cox Proportional Hazards Model]
The formula for the Cox Proportional Hazards Model is given by

$$h(t, \boldsymbol{X}) = h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_j\right) \tag{4}$$

with $\boldsymbol{X}$ being the collection of time-independent explanatory variables (covariates), β_j are the coefficients representing the effect of X_j on the hazard and p is the amount of covariates.

$$\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_p \end{pmatrix}$$

The Cox model (4) introduces the baseline hazard function $h_0(t)$. It is unspecified function, which represents the hazard function $h(t, \boldsymbol{X})$ without considering any of the explanatory variables, that is, when $X_1 = X_2 = \cdots = X_p = 0$. Since the baseline hazard function is not specified, the Cox PH model is referred to as semi-parametric model, which also provides the flexibility to the model by allowing for less restrictions of the model [Kleinbaum and Klein, 2005]. The key benefit of the Cox proportional hazards model is the assessment the effect of several explanatory variables on the survival time.

We use the hazard ratio (HR), in order to compare the hazards $h(t, \boldsymbol{X}^*)$ and $h(t, \boldsymbol{X})$ for $\boldsymbol{X}^* = (X_1^* X_2^* \cdots X_p^*)$ being the collection of covariates for the second individual. In the HR, the unspecified baseline hazard function is reduced, that is:

$$\begin{aligned} HR &= \frac{h_0(t)e^{\sum_{j=1}^{p}\beta_j X_j^*}}{h_0(t)e^{\sum_{j=1}^{p}\beta_j X_j}} \\ &= e^{\sum_{j=1}^{p}\beta_j (X_j^* - X_j)} \end{aligned} \tag{5}$$

Therefore the hazard ratio allows us to compare the effects of different groups of covariates on the individuals hazard. The interpretation of HR is straightforward: if $HR > 1$, then the risk of the event of interest occurring is higher for individual with $\boldsymbol{X}^*$ set of covariates than for the individual with $\boldsymbol{X}$ set of covariates; if $HR < 1$, then the hazard is higher for the individual with $\boldsymbol{X}$ collection; if $HR = 1$, then the risks are the same. That is, if $HR = 10$, then the risk of the event happening is 10 times higher for the individual with set $\boldsymbol{X}^*$, than for the other individual; if $HR = 1/10$, then 10 times lower.

3.1.4 Estimation of coefficients

In this subsection we briefly elaborate on how the regression coefficients β_j are obtained for the explanatory variables X_j in the Cox proportional hazards model [Kleinbaum and Klein, 2005].

Recall, that the likelihood function $L(\boldsymbol{\beta})$ describes the joint probability of obtaining the observed data $\boldsymbol{X}$ as the function of coefficients $\boldsymbol{\beta}$.

Definition 1.4. [Likelihood function]
The (partial) likelihood function of Cox Proportional Hazards model is given by

$$\begin{aligned} L(\boldsymbol{\beta}) &= L_1 \times L_2 \times \ldots \times L_n \\ &= \prod_{i=1}^{n} L_i \end{aligned} \tag{6}$$

where L_i is the likelihood function at individual $i = 1, \ldots, n$, given that the individual survived up to the event times (failure times).

The individual likelihood L_i at t_i failure time is defined as

$$\begin{aligned} L_i(\boldsymbol{\beta}) &= \frac{\exp(\boldsymbol{X_i\beta})}{\sum_{l \in R(t_i)} \exp(\boldsymbol{X_l\beta})} \\ &= \frac{\exp(\sum_{j=1}^{p} \beta_j X_{ij})}{\sum_{l \in R(t_i)} \exp(\sum_{j=1}^{p} \beta_j X_{lj})} \end{aligned} \tag{7}$$

where the risk set $R(t_i)$ is the set of the individuals $l = 1, \ldots, n$ for which the event did not yet happen at $t_i = 1, \ldots, k$ failure times. We note that we adjust the definition given in [Kleinbaum and Klein, 2005], for accounting for the individuals having the same failure time and the definition of the set of covariates for the corresponding individual.

That is, $\boldsymbol{X_i}$ is the set/vector of p covariates for the individual i

$$\boldsymbol{X_i} = \begin{pmatrix} X_{i1} & X_{i2} & \cdots & X_{ip} \end{pmatrix}$$

Thus, in (7), X_{ij} is the entry for the vector $\boldsymbol{X_i}$ for $j = 1, \ldots, p$ covariate. Similarly we define X_{lj} for the individuals l in the risk set.

For the Cox regression, $L(\boldsymbol{\beta})$ is called a partial likelihood (PL) function, since $L(\boldsymbol{\beta})$ does not account for the baseline hazard function.

The estimates of coefficients β_j are obtained through maximization of the log-likelihood function $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta})$ and are called the maximum likelihood estimates (MLE), denoted as $\hat{\beta}_j$. We maximize $l(\boldsymbol{\beta})$ by solving the score equations, that is, taking the partial derivatives of $l(\boldsymbol{\beta})$ wrt β_j and setting the equations to 0: $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = 0$, $j = 1 \ldots p$, where p is number of covariates in $\boldsymbol{X}$ [Casella and Berger, 2002].

There are several extensions of the Cox proportional model, designed and developed to be implemented for different data structures. In the next section we present such extension and motivate our choice of the model.

3.2 Model development

Previously we provided the description of our data collected, specifically elaborating on the variables description and their further categorization. We recall, that the explanatory variable Distributor is the categorical variable, consisting of 57 studios which produced the films, being the "initial" stage of the filmmaking process. Thus, the movies are nested within the Distributors, and the movies from the same distributor are likely to share unobserved in our research common characteristics, thus not being independent of each other. Therefore we introduce the multilevel structure for our data, placing the Distributors on the highest (second) level of the hierarchy and the individual-specific covariates (Opening Gross, Genres, Percentage of Open Theatres to Maximum Theatres, Rating, and Running Time) on the lower (first) level[6].

We therefore argue, that the (standard) Cox proportional hazards model might not fully capture our hierarchical by structure data and due to the unobserved characteristics and large amount of Distributors. In this section by introducing the extension to the ordinary Cox PH model, namely the Two level Cox Regression Model with Mixed Effects, we can take into account the variability across movies that belong to the different distributors.

3.2.1 Mixed effects

We call the model mixed, when it contains both, fixed effects and random effects. When we refer to all of the regression coefficients of the model, we therefore use the term mixed effects. Mixed models are beneficial over, for example, the standard Cox Proportional Hazards model, if we have the nested data [Austin, 2017]. In such case, the random effects modify the baseline hazards function (4) to account for the within-cluster correlation, while the fixed effects are the same as for the Cox PH model. We further introduce the estimation of the fixed and random effects coefficients.

3.2.2 Shared Frailty model

The frailty model denotes the survival model that includes the random effects and is the special case of the mixed effects model[7] [Austin, 2017]. For our extension of the Cox proportional hazards model, we incorporate the frailty, which accounts for the variability in the model due to the unobserved factors affecting the model's outcome variable (the survival time). There are two types of frailty model: Unshared frailty model and Shared frailty model [Kleinbaum and Klein, 2005].

In the unshared frailty model the frailty is distributed independently among the individuals. That is, each movie has the unique, unobserved in the collected data, factors which affect their individual survival.

[6] In some literature for the multilevel survival analysis structure, the survival time is placed on the lowest (zero) level of the hierarchy

[7] We have one random intercept for the Distributors. The "multiple random effects" models include 3-level or higher structures

In the shared frailty model, the groups (clusters) share the same frailty, that is, the shared frailty accounts for the individuals from the same cluster sharing the unobserved factors affecting their survival time. That is, within cluster correlation for all movies. In our research we assume that the movies from the same distributors share the common unobserved in our data factors affecting their survival time. Thus, we propose the shared frailty model in our research, where the random effects measure the increase or decrease in hazard, depending on the specific Distributor indicator for movies.

The Cox shared frailty model is of the general form:

$$h_i(t \mid \alpha_k) = \alpha_k h_0(t) \exp(\boldsymbol{X}_i \boldsymbol{\beta}) \tag{8}$$

for $i = 1, 2, \ldots, n_k$, where n_k is a total amount of individuals in the k^{th} cluster; α_k is the frailty shared in group k, representing the frailty of a specific kth Distributor.

Thus, the model (8) introduces shared (cluster-specific) frailty, which has the multiplicative effect on the baseline hazard function $h_0(t)$, modifying it for corresponding clusters for the within-cluster correlation [Austin, 2017].

3.2.3 Two-Level Cox Regression Model with Mixed Effects

In order to develop the understanding of the general formula (8) by means of our data, we introduce the following definition.

Definition 1.5. [Two-Level Cox Regression Mixed Effects Model]
The formula for the Two-Level Cox Regression Mixed Effects Model is given by

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b}\right) \tag{9}$$

where $i = 1, ..., n$ are the individuals/movies considered in the model. Note that $\exp(\boldsymbol{Z_i}\boldsymbol{b})$ is thus the frailty term in (9).

In the model (9), we have: The design (regression) matrix $\boldsymbol{X}$ of dimension $n \times p$ for fixed effects, a matrix of values of explanatory variables, where n is the amount of movies, p is the amount of covariates (see the categorization of variables), the row corresponds to a movies and each column to the covariate indicator; $\boldsymbol{\beta}$ is the vector of the fixed effects coefficients, representing the effect on hazard for different individual-level covariates; $\boldsymbol{b}$ is the vector of random effects coefficients.

The design matrix $\boldsymbol{Z}$ of dimension $n \times k$, where n is the amount of movies, k is the amount of distributors, for random effects is defined as

$$\boldsymbol{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{bmatrix}$$

Since each distributor is a cluster, we define the entries as binary variables, each entry z_{ij} for $j = 1, ..., k$ takes the binary value, that is, $z_{ij} = 1$ if the movie i belongs to the distributor j, $z_{ij} = 0$ otherwise.

Given the structure of the design matrices, we observe the complexity which we would have had if treating distributor's coefficients as fixed effects, in addition to the unobserved common characteristics. That is, recalling the categorization of other covariates in the Data section, the amount of Distributors exceeds in size (57 distributor's categories) any other categorization of the variables, which increases the complexity of the estimation procedure of the fixed effect coefficients, if treating Distributors as individual-level covariate.

In the model (9), the random effects vector $\boldsymbol{b}$ are assumed to follow some distribution, which depends on the unknown parameter. We assume that the random effects follow the Multivariate normal distribution:

$$\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\theta)) \tag{10}$$

Where for the shared frailty, b_j are independent and identically distributed (i.i.d), the covariance matrix $\boldsymbol{\Sigma}(\theta)$ reduces to $\boldsymbol{\Sigma} = \theta \boldsymbol{I}$, $\boldsymbol{I}$ is the identity matrix and θ is the variance of the distribution of b_j. This means that we assume that the random effects associated with each Distributor follow the same distribution, while the effect on the hazard for every Distributor differs, since the estimated random effects coefficients will differ. Thus, we introduce the random effects for the within-cluster correlation of movies under the same distributor.

That is, for the cluster $j = 1, \ldots, k$, b_j has the Gaussian distribution with mean 0, variance θ, and the probability density function (pdf) for b_j is

$$f(b_j) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{b_j^2}{2\theta}\right) \tag{11}$$

For $\boldsymbol{b}$, the vector of random effects, we recall the pdf of Multivariate normal distribution with mean $\mathbf{0}$ and the positive definite covariance matrix $\boldsymbol{\Sigma}(\theta)$:

$$f(\boldsymbol{b}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}(\theta)|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{b}'\boldsymbol{\Sigma}(\theta)^{-1}\boldsymbol{b}\right) \tag{12}$$

where k - is the number of random effects in $\boldsymbol{b}$ [Therneau, 2022].

Note, that the frailty (8) will thus have the log-normal distribution.

The choice of the distribution is done following [Therneau and Grambsch, 2000] theoretical approach, and as in [Austin, 2017], the authors indicate that more research is needed in order to validate the distribution choice.

3.3 Inference for mixed effects coefficients

In this section we describe the estimation of regression coefficients in the Two-level Cox regression model with Mixed effects.

3.3.1 Likelihood function of the Two-Level Cox Regression Mixed Effects Model

Definition 1.6. [Likelihood function]
The likelihood function for the Cox mixed effects model is defined as

$$L(h_0(t), \boldsymbol{\beta}, \theta) = \int_{\boldsymbol{b}} \prod_{i=1}^{n} h_i(t|\boldsymbol{b})^{\delta_i} S_i(t|\boldsymbol{b})\, g(\boldsymbol{b}; \boldsymbol{\Sigma}(\theta))\, d\boldsymbol{b} \tag{13}$$

where $\delta_i = I_{T_i \leq C_i}$ is the indicator function for event times T_i and censoring times C_i for individuals $i = 1, ..., n$. $\delta_i = 1$ indicates that the event is observed within the study period (not censored), or the event time is censored, $\delta_i = 0$ [Ripatti and Palmgren, 2000]. For our project, the event happened for all individuals. Recalling the covered basics for the survival analysis, $\prod_{i=1}^{n}[h_i(t|\boldsymbol{b})^{\delta_i} S_i(t|\boldsymbol{b})]$ is the joint likelihood for the individuals i at time t, given the random effects.

The likelihood function for mixed effects Cox model introduces the penalty function $g(\boldsymbol{b}; \boldsymbol{\Sigma}(\theta))$. The penalty function aligns with the distribution of the random effects discussed previously, in (13) multiplication by penalty penalizes the extreme values of $\boldsymbol{b}$ [Ripatti and Palmgren, 2000]. The penalty function is defined as the pdf of random effects, $\boldsymbol{b} \sim g(\boldsymbol{b}; \boldsymbol{\Sigma}(\theta))$, with mean $\mathbf{0}$, that is:

$$g(\boldsymbol{b}; \boldsymbol{\Sigma}(\theta)) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}(\theta)|^{1/2}} \exp\left(-\frac{1}{2} \boldsymbol{b}' \boldsymbol{\Sigma}(\theta)^{-1} \boldsymbol{b}\right) \tag{14}$$

where k is the amount of random effects.

We note that (13) is the marginal likelihood function. The random effects are unknown, therefore the idea is to integrate them out in order to average over the distribution of random effects [Fitzmaurice et al., 2011] and to estimate θ.

Let $\boldsymbol{\eta_i} = \boldsymbol{X_i \beta} + \boldsymbol{Z_i b}$. Now, recalling the formulas (9) and (3), we can rewrite the marginal likelihood function as:

$$L(h_0(t), \boldsymbol{\beta}, \theta) = \int_{\boldsymbol{b}} \prod_{i=1}^{n} [h_0(t) \exp(\boldsymbol{\eta_i})]^{\delta_i} \exp\left[-\int_0^t h_0(u)\, du \exp(\boldsymbol{\eta_i})\right] g(\boldsymbol{b}; \boldsymbol{\Sigma}(\theta))\, d\boldsymbol{b} \tag{15}$$

where for rewriting the survival function $S_i(t|\boldsymbol{b})$ we integrate out $\exp(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b})$ since we do not have time-dependent coefficients in our model.

We also note that in the term $\exp\left[-\int_0^t h_0(u)\, du \exp(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b})\right]$ integrating over the unspecified baseline hazard function $h_0(u)$ makes the model (15) the "full" likelihood function.

In the research of [Ripatti and Palmgren, 2000], the authors start with the full marginal likelihood function (13), approximate the integral and reduce the terms in the resulting approximation, presenting the (penalized) partial likelihood function to maximize, that is, deriving the score functions for $\boldsymbol{\beta}$ and $\boldsymbol{b}$. In our work we proceed with the different approach for deriving the score functions, starting with the partial likelihood function, following and filling in the steps in the [Therneau and Grambsch, 2000] and [Therneau, 2022][8] literature, written by the author of the "survival" R package Terry Therneau. Our motivation is in better understanding of the derivations and presentation of the likelihood functions. We also believe that starting from the partial likelihood function helps with better understanding of the estimation procedure of the Cox models, allowing for the step-wise approach for defining each intermediate equation. We note that the reader is also advised to consult the approach of [Ripatti and Palmgren, 2000].

3.3.2 Partial Likelihood function

Definition 1.7. [Partial Likelihood function]
The formula for the partial likelihood (PL) function for Cox model with mixed effects is given by

$$PL(\boldsymbol{\beta}, \boldsymbol{b}) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{Y_i(t) \exp(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b})}{\sum_j Y_j(t) \exp(\boldsymbol{X_j \beta} + \boldsymbol{Z_j b})} \right\}^{dN_i(t)} \tag{16}$$

We note that the form of (16) can be seen as the ordinary Cox partial likelihood, by letting $\boldsymbol{\eta_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b}$.

For the partial likelihood function (16), instead of the risk set defined previously, we introduce the functions as in [Therneau and Grambsch, 2000], where the definitions were generalized to account for different events occurring and for the counting process:

$$\begin{aligned} N_i(t) &= \text{ number of observed events for } i \text{ in } [0, t], \\ Y_i(t) &= I\{T_i \geq t\}. \end{aligned}$$

where $Y_i(t)$ is the risk set indicator function; $N_i(t)$ indicates the event times.

Again, we recall that for our research we have no censored events and we do not have the recurrent events for the individuals i in our data. However, we keep the definition (16), for the potential use in further researches, where the censoring of the data is present, or the events are recurring[9].

That is, the indicator functions are be interpreted as:

$$dN_i(t) = \begin{cases} 1 & \text{if for an individual } i \text{ the event happened at } t, \\ 0 & \text{otherwise.} \end{cases}$$

$$Y_i(t) = \begin{cases} 1 & \text{if the individual } i \text{ is at risk at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

That is, dN_i is the indicator of the event times T_i.

[8] The documentation for the estimation of regression coefficients in R for the Cox Mixed Effects model

[9] In the case of the movie data, for example, censoring would occur for the movies still running in theatres at the time of the research conducted; and recurrent events (multiple events) could happen for films re-released during the timeframe of the data collected

3.3.3 Log-likelihood function of the Two-Level Cox Regression Mixed effects Model

Given the definitions of the indicator function above, we can derive the formula for the Log-likelihood function of (16). By deriving the equation we can understand the indicator functions better, and the log-form is needed in order to derive the score function.

$$\begin{aligned}
\log PL(\boldsymbol{\beta}, \boldsymbol{b}) &= \log \left(\prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{Y_i(t) \exp(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b})}{\sum_j Y_j(t) \exp(\boldsymbol{X_j \beta} + \boldsymbol{Z_j b})} \right\}^{dN_i(t)} \right) \\
&= \sum_{i=1}^{n} \sum_{t \geq 0} dN_i(t) \cdot \log \left\{ \frac{Y_i(t) \exp(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b})}{\sum_j Y_j(t) \exp(\boldsymbol{X_j \beta} + \boldsymbol{Z_j b})} \right\} \\
&= \sum_{\substack{i=1 \\ Y_i(t)=1}}^{n} \sum_{t \geq 0} dN_i(t) \cdot \left(\log \exp(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b}) - \log \sum_j Y_j(t) \exp(\boldsymbol{X_j \beta} + \boldsymbol{Z_j b}) \right) \\
&= \sum_{\substack{i=1 \\ Y_i(t)=1}}^{n} \sum_{t \geq 0} dN_i(t) \cdot \left(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b} - \log \sum_j Y_j(t) \exp(\boldsymbol{X_j \beta} + \boldsymbol{Z_j b}) \right) \\
&= \sum_{\substack{i=1 \\ Y_i(t)=1}}^{n} \int_0^{\infty} \left(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b} - \log \sum_j Y_j(t) \exp(\boldsymbol{X_j \beta} + \boldsymbol{Z_j b}) \right) dN_i(t) \\
&= \sum_{i=1}^{n} \int_0^{\infty} \left(Y_i(t)(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b}) - \log \sum_j Y_j(t) \exp(\boldsymbol{X_j \beta} + \boldsymbol{Z_j b}) \right) dN_i(t)
\end{aligned} \tag{17}$$

Finding the MLE estimate for the parameters would be done through maximizing the Partial log-likelihood function, the solution for estimating $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$, found through setting score function equal to 0. However, since we additionally to fixed effects have the random effects, we would need to maximize a different likelihood function, which accounts for the clusters variations, in order to get the solution vector $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}})$. Therefore, next we introduce the Penalized Partial likelihood (PPL) function, which concept corresponds to the approach of [Ripatti and Palmgren, 2000].

3.3.4 Penalized Partial (log-)Likelihood

We define The Penalized Partial likelihood (PPL)[10] as:

$$PPL = PLe^{-\frac{1}{2}\boldsymbol{b}'\boldsymbol{\Sigma}(\theta)^{-1}\boldsymbol{b}} \tag{18}$$

Remark: the penalty term in the (18) is not the "full" penalty function defined for the full marginal likelihood. This is clear by following the approach of [Ripatti and Palmgren, 2000], however, their derivation is not detailed, as we have already mentioned, the PPL function

[10] In our main literature used, the authors define the log PPL (LPPL) directly, in our work we provide the intermediate steps for completeness

was introduced by reducing some terms of the approximation to the full likelihood function. However, the penalty term serves the same purpose, that is regularizing the random effects, by penalizing the extreme values of the random effects coefficients as indicated in both, [Therneau and Grambsch, 2000] and [Ripatti and Palmgren, 2000]. Based on the different distribution choices, the penalty term differs, and in (18) the penalty for the multivariate distribution is given.

That is, for the cluster j the PPL is defined as [Therneau and Grambsch, 2000]:

$$LPPL = LPL - \sum_{j=1}^{k} \frac{b_j^2}{2\theta} \tag{19}$$

We can now provide the PPL in the vector-matrix form.

Definition 1.8. [Penalized Partial Log-Likelihood]
The Penalized Partial Log-Likelihood (LPPL) for the shared frailty is given by

$$\begin{aligned} LPPL &= LPL - \sum_{j=1}^{k} \frac{b_j^2}{2\theta} \\ &= LPL - \frac{1}{2\theta} \sum_{j=1}^{k} b_j^2 \\ &= LPL - \frac{1}{2\theta} \boldsymbol{b}' \boldsymbol{I} \boldsymbol{b} \\ &= LPL - \frac{1}{2} \boldsymbol{b}' \left(\frac{\boldsymbol{I}}{\theta} \right) \boldsymbol{b} \\ &= LPL - \frac{1}{2} \boldsymbol{b}' \boldsymbol{\Sigma}(\theta)^{-1} \boldsymbol{b} \\ &= \sum_{\substack{i=1 \\ Y_i(t)=1}}^{n} \sum_{t \geq 0} dN_i(t) \cdot \left(\boldsymbol{X_i \beta} + \boldsymbol{Z_i b} - \log \sum_j Y_j(t) \exp(\boldsymbol{X_j \beta} + \boldsymbol{Z_j b}) \right) - \frac{1}{2} \boldsymbol{b}' \boldsymbol{\Sigma}(\theta)^{-1} \boldsymbol{b} \end{aligned} \tag{20}$$

where $\boldsymbol{I}$ is the identity matrix; $-\frac{1}{2}\boldsymbol{b}'\Sigma(\theta)^{-1}\boldsymbol{b}$ is the penalty term in the quadratic form; we set $\boldsymbol{\Sigma} = \theta \boldsymbol{I}$ for the shared frailty; θ is the variance of the random effect. In the last step we plug in the (17) equation.

We also recognize that in [Ripatti and Palmgren, 2000] the authors derived the similar formula for the LPPL following the full marginal likelihood approach.

3.3.5 Estimation equations for $\hat{\beta}$ and $\hat{b}$

For estimating the fixed effects $\hat{\boldsymbol{\beta}}(\theta)$ and random effects $\hat{\boldsymbol{b}}(\theta)$ we maximize the log Penalized partial likelihood (20), [Therneau and Grambsch, 2000]. Thus, we differentiate the LPPL for the expression for the score vector functions, [Ripatti and Palmgren, 2000] using the basic differentiation rules.

$$\frac{\partial LPPL(\boldsymbol{\beta},\boldsymbol{b})}{\partial\boldsymbol{\beta}} = \sum_{\substack{i=1\\Y_i(t)=1}}^{n} \sum_{t\geq 0} dN_i(t)\cdot\left(\boldsymbol{X_i} - \frac{\boldsymbol{X_i}\exp(\boldsymbol{X_i\beta}+\boldsymbol{Z_i b})}{\sum_j Y_j(t)\exp(\boldsymbol{X_j\beta}+\boldsymbol{Z_j b})}\right) \tag{21}$$

Now, setting the above score vector function of dimension $p\times 1$ to 0, we get the estimating equation for $\hat{\boldsymbol{\beta}}(\theta)$. Note that the penalty term vanishes.

For the estimating equation for $\boldsymbol{b}$, we have the following $k\times 1$ score vector function:

$$\frac{\partial LPPL(\boldsymbol{\beta},\boldsymbol{b})}{\partial\boldsymbol{b}} = \sum_{\substack{i=1\\Y_i(t)=1}}^{n} \sum_{t\geq 0} dN_i(t)\cdot\left(\boldsymbol{Z_i} - \frac{\boldsymbol{Z_i}\exp(\boldsymbol{X_i\beta}+\boldsymbol{Z_i b})}{\sum_j Y_j(t)\exp(\boldsymbol{X_j\beta}+\boldsymbol{Z_j b})}\right) - \boldsymbol{\Sigma}(\theta)^{-1}\boldsymbol{b} \tag{22}$$

Again, setting the above score function to 0, we can derive the estimator $\hat{\boldsymbol{b}}(\theta)$. For deriving the estimating equations for $(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))$, we alternate between solving for the two score functions [Ripatti and Palmgren, 2000], using the iterative Newton-Raphson [Therneau and Grambsch, 2000].

That is, we iteratively compute

$$\hat{\boldsymbol{\beta}}^{(n+1)} = \hat{\boldsymbol{\beta}}^{(n)} + \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}^{(n)})\boldsymbol{U}(\hat{\boldsymbol{\beta}}^{(n)}) \tag{23}$$

where $\boldsymbol{U}(\boldsymbol{\beta})$ is the score function (21); $\mathcal{I}(\boldsymbol{\beta})$ is the information matrix, as described in [Therneau and Grambsch, 2000], and we follow the same procedure for the random effects estimating equation.

The alternative estimation procedures and their efficiency are described in [Zhang, 2006], in the dissertation written under the supervision of Terry Therneau.

The difficulty with estimation arises with the unknown θ, since the estimating equations for mixed effects coefficients depend on the variance of random effects, $(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))$. Therefore we address the estimation for $\hat{\theta}$ in the next section and the general algorithm is described at the end of the section.

3.3.6 Integrated Partial Likelihood

Since we are interested in the likelihood function depending on the parameters $\boldsymbol{\beta},\boldsymbol{b},\theta$, the idea is to present the Integrated Partial likelihood (IPL) where we integrate over the random effects in the penalized partial likelihood model.

The IPL is the marginal likelihood, where we integrate out the random effect coefficients. Now we see how the approach described at the beginning of the section, for the full likelihood function, relates to ours. That is, due to the additional complexity of the model, the inclusion of frailty, by integrating we get the likelihood function $IPL(\boldsymbol{\beta},\theta)$ which depends on the $\boldsymbol{\beta},\theta$, where we recall that θ is the variance for the distribution of b_j. Therefore the idea of the integration is to estimate the covariance of $\boldsymbol{b}$, $\theta \boldsymbol{I}$ [Fitzmaurice et al., 2011].

Definition 1.9. [Integrated Partial Likelihood
The Integrated Partial Likelihood (IPL) is defined as

$$IPL(\boldsymbol{\beta}, \theta) = \int PL(\boldsymbol{\beta}, \boldsymbol{b}, \theta) g(\boldsymbol{b}; \boldsymbol{\Sigma}(\theta)) d\boldsymbol{b} \tag{24}$$

Having the IPL likelihood function, we can now maximize it to get the ML estimate for θ. However, since the integral is intractable, that is, difficult to compute directly, we approximate the integral using the Laplace approximation.

For the reader, before proceeding with the approximation, we make the short note about the similarity of the IPL with the marginal likelihood function (the denominator) in the Bayes equation [Dobson and Barnett, 2018].

Recall the Bayes equation,

$$P(\theta|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\theta)P(\theta)}{P(\boldsymbol{y})} \tag{25}$$

where $\boldsymbol{y}$ is the given data and θ is the unknown parameter; $P(\theta|\boldsymbol{y})$ is posterior probability of θ; $P(\boldsymbol{y}|\theta)$ is the likelihood function; $P(\theta)$ is prior; $P(\boldsymbol{y})$ is the probability of the data $\boldsymbol{y}$.

For the continuous parameter space, we can rewrite the probability of the data $\boldsymbol{y}$ through the parameter θ:

$$P(\boldsymbol{y}) = \int P(\boldsymbol{y}|\theta)P(\theta)d\theta \tag{26}$$

where we integrate over the unknown θ.

Before the approximation, we provide the relationship of the model (24) with PPL (18):

$$\begin{aligned} IPL(\boldsymbol{\beta}, \theta) &= \int PL(\boldsymbol{\beta}, \boldsymbol{b}, \theta) g(\boldsymbol{b}; \theta) d\boldsymbol{b} \\ &= (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}(\theta)|^{-\frac{1}{2}} \int PL(\boldsymbol{\beta}, \mathbf{b}) e^{-\frac{1}{2}\boldsymbol{b}'\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{b}} d\boldsymbol{b} \\ &= (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}(\theta)|^{-\frac{1}{2}} \int e^{LPL-\frac{1}{2}\boldsymbol{b}'\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{b}} d\boldsymbol{b} \\ &= (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}(\theta)|^{-\frac{1}{2}} \int e^{LPPL} d\boldsymbol{b} \\ &= (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}(\theta)|^{-\frac{1}{2}} \int e^{\log PPL} d\boldsymbol{b} \\ &= (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}(\theta)|^{-\frac{1}{2}} \int PPL(\boldsymbol{\beta}, \boldsymbol{b}, \theta) d\boldsymbol{b} \end{aligned} \tag{27}$$

Now we can use Laplace approximation to the integral in (27). The general idea is described in [Therneau, 2022], and in our project we present the stepwise approximation, in order to fill in the gaps and for the better understanding of the estimation procedure.

Following the explanation of the idea of Laplace approximation in [MacKay, 2003], we first use Taylor expansion for $\log(PL(\boldsymbol{\beta},\boldsymbol{b})e^{-\frac{1}{2}\boldsymbol{b}'\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{b}})$ around the solution vectors $(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))$:

$$LPL(\boldsymbol{\beta},\theta)-\frac{1}{2}\boldsymbol{b}'\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{b}\approx LPPL(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}(\theta),\boldsymbol{b}-\hat{\boldsymbol{b}}(\theta))'\boldsymbol{H}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}(\theta),\boldsymbol{b}-\hat{\boldsymbol{b}}(\theta))+\ldots \tag{28}$$

Where Hessian $\boldsymbol{H}$ is the negative of second derivatives of LPPL evaluated at $(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))$, that is:

$$\boldsymbol{H}=-\left.\frac{\partial^2 LPPL(\boldsymbol{\beta},\boldsymbol{b})}{\partial\boldsymbol{b}\partial\boldsymbol{b}'}\right|_{(\boldsymbol{\beta},\boldsymbol{b})=(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))} \tag{29}$$

Now we can approximate the integral as

$$\begin{aligned}\int PL(\boldsymbol{\beta},\boldsymbol{b})e^{-\frac{1}{2}\boldsymbol{b}'\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{b}}d\boldsymbol{b}&\approx PL(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))e^{-\frac{1}{2}\hat{\boldsymbol{b}}'\boldsymbol{\Sigma}^{-1}(\theta)\hat{\boldsymbol{b}}}\frac{1}{\sqrt{|\frac{1}{2\pi}\boldsymbol{H}|}}\\&=PL(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))e^{-\frac{1}{2}\hat{\boldsymbol{b}}'\boldsymbol{\Sigma}^{-1}(\theta)\hat{\boldsymbol{b}}}(2\pi)^{\frac{k}{2}}\cdot|\boldsymbol{H}|^{-\frac{1}{2}}\end{aligned} \tag{30}$$

where k is the amount of random effects, that is, the dimension of $k\times 1$ vector $\boldsymbol{b}$.

Now, plugging the approximation to the IPL, we have

$$\begin{aligned}IPL&=(2\pi)^{-\frac{k}{2}}\,|\boldsymbol{\Sigma}(\theta)|^{-\frac{1}{2}}\,PL(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))e^{-\frac{1}{2}\hat{\boldsymbol{b}}'\boldsymbol{\Sigma}^{-1}(\theta)\hat{\boldsymbol{b}}}(2\pi)^{\frac{k}{2}}\cdot|\boldsymbol{H}|^{-\frac{1}{2}}\\&=|\boldsymbol{\Sigma}(\theta)|^{-\frac{1}{2}}\,PL(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))e^{-\frac{1}{2}\hat{\boldsymbol{b}}'\boldsymbol{\Sigma}^{-1}(\theta)\hat{\boldsymbol{b}}}|\boldsymbol{H}|^{-\frac{1}{2}}\end{aligned} \tag{31}$$

The log form of the Integrated Partial Likelihood is given by:

$$\log(IPL)=-\frac{1}{2}\log|\boldsymbol{\Sigma}(\theta)|+LPL(\hat{\boldsymbol{\beta}}(\theta),\hat{\boldsymbol{b}}(\theta))-\frac{1}{2}\hat{\boldsymbol{b}}'\boldsymbol{\Sigma}^{-1}(\theta)\hat{\boldsymbol{b}}-\frac{1}{2}\log|\boldsymbol{H}| \tag{32}$$

Finally, the estimating equation for θ is derived and described in [Ripatti and Palmgren, 2000] and in [Therneau and Grambsch, 2000], therefore we do not replicate their derivation, but provide the formula for the derived variance $\hat{\theta}$ for completeness. For the shared frailty, we have

$$\hat{\theta}=\frac{\hat{\boldsymbol{b}}'\hat{\boldsymbol{b}}+trace(\boldsymbol{H})^{-1}}{k} \tag{33}$$

where k is the amount of random effects.

As the final note, later we use the coxme package developed by Terry Therneau, whom statistical texts we followed throughout this section, that is, the general procedure of estimating the mixed effects coefficients for the coxme function is the same, as we described, but computationally optimized for the implementation in R. That is:

1. For each trial $\boldsymbol{\Sigma}(\theta)$ calculate $\boldsymbol{\Sigma}^{-1}(\theta)$
2. Solve estimating equations derived from PPL for $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}})$ with iterative Newton-Raphson

3. Compute log IPL

Repeating the algorithm will find the actual θ that maximizes log IPL. For further details consult [Therneau, 2022].

3.4 Model Comparison and Validation

3.4.1 AIC and BIC

When it comes to the model selection, we consider AIC (Akaike information criterion) and BIC (Bayesian information criterion) values of the models. The AIC and BIC are the measures of the information loss by fitting the model of interest. Therefore the lower values correspond to the better model fit. That is, the value of AIC illustrates the compromise between the goodness of fit and the simplicity of the model [Huang, 2022]. For the AIC the formula we use is

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{34}$$

And for the BIC we have

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L}) \tag{35}$$

where n is the number of observations, k is the number of estimated parameters and $\ln(\hat{L})$ is the log-likelihood of the model with the ML estimated parameters.

The BIC also penalizes the model for over-fitting, but with the larger penalty [Huang, 2022]. In (35), BIC depends on the number of observations, that is, the sample size, thus the penalty of $\ln(n)k$ is larger for bigger[11] sample size, than the penalty $2k$ in AIC definition (34).

We compare the models by considering the differences between the corresponding AIC values. The model with the lower AIC is considered to be of better fit and is preferred over the model with the higher AIC. Similarly for the BIC values comparison. Below is the summary of the interpretation of the differences [Burnham and Anderson, 2004], [Berchtold, 2010].

AIC Difference (Δ_{AIC})	**Interpretation**
$\Delta_{\text{AIC}} \leq 2$	Insignificant: No model preference
$4 \leq \Delta_{\text{AIC}} \leq 7$	Significant: The model with lower AIC is considered
$10 < \Delta_{\text{AIC}}$	Strong support of the lower AIC model

Table 7: Interpretation of Differences in AIC

[11] that is, more than 7 observations, since $\exp(2) \approx 7.39$

BIC Difference (Δ_{BIC})	**Interpretation**
$\Delta_{\text{BIC}} \leq 2$	Insignificant: No model preference
$2 < \Delta_{\text{BIC}} \leq 5$	Somewhat Significant: There is the preference of the model with lower BIC
$5 < \Delta_{\text{BIC}} \leq 10$	Significant: Strong choice of the lower BIC model
$10 < \Delta_{\text{BIC}}$	Very strong support of the lower BIC model

Table 8: Interpretation of Differences in BIC

3.4.2 Assessing the Normality Assumption

Previously we mentioned that the normality assumption, that is, the normal distribution choice of random effects (log-normal for frailty) can be considered as the preferable approach. However, the limitation of the choice between the distributions, for example, between gamma and log-normal is present [Therneau and Grambsch, 2000], [Austin, 2017]. The further research is required, as the authors stated, while in our project the choice of the multivariate distribution was made for the shared frailty model, following T. Therneau preference. As [Austin, 2017] mentioned regarding the gamma distributed frailty, methods for comparison with different than gamma distribution are limited, not allowing for comparison, but for the assessment of gamma distribution. Therefore in our thesis we will assess the normality assumption, rather than compare different distributions.

The normal distribution of random effects can be assessed using the Q-Q plot (quantile to quantile plot) [Sun et al., 2023]. In the Q-Q plot we compare the empirical distribution, that is, the sample distribution of the random effects coefficients, with the theoretical normal distribution in the quantile-quantile coordinates. The construction of the plot can be described as following [Dodge, 2008]:

We first arrange the random effect coefficient in the increasing order, that is:

$\hat{b}_{(1)} \leq \hat{b}_{(2)} \leq ... \leq \hat{b}_{(57)}$, where the subscript indicates the sorted position. Then we mark the values of random effects $\hat{b}_{(i)}$ on the y-axis on our plot. Next we find the positions measured in the quantiles of the standard normal distribution. That is, for each $\hat{b}_{(i)}$ the corresponding quantile is $q_i = \frac{i}{n+1}$ of the standard normal distribution[12] and mark the values on the x-axis. In the plot, if the points are on the $x = y$ line, then this confirms the choice of the normal distribution. Otherwise, if the random effects significantly deviate from the line of theoretical normal distribution, the normality assumption is violated.

[12] In the documentation for R functions for q-q plot (qqPlot), we found that the formula for quantiles is sometimes adjusted for better plot positions and the correspondence to the empirical cumulative distribution function (empirical cdf) $\mathrm{F}(x_{(i)}) = q_i$ is described [Rteam, 2019].

3.4.3 Harrell's C-index

In order to validate our Cox mixed effects model, we perform the cross-validation on our data. The k-fold cross-validation can be described as following [Hastie et al., 2009]: We divide our dataset into k folds (k subsets of data), approximately equal sized. Then we use k-1 folds for training, that is, fitting the model on the Train data and the kth fold (not used in training) is the Test data (the validation kth fold). On the Test data we predict the estimates by using the Train data model fit results. For the accurate prediction, we repeat the prediction by changing the Test data for every $i = 1, \dots, k$ fold. Then the assessment of the accuracy of the prediction is done, as we describe next.

There are several cross-validation approaches for assessing the Cox proportional hazards model. However, for the Cox mixed effects models, currently there is no method of validation, which is considered to be standard for the model assessment. When we researched upon the methods, only a few papers discussed the measures for cross-validation for the extended Cox models. [Dai and Breheny, 2019] proposed a few possible adaptation of cross-validation methods to the penalized Cox model. In this study, the Concordance index (C-index) was used as one of the measures of the ability for the model to accurately predict the hazards for different individuals. Upon further search, there are several adaptations to the C-index definitions, accounting for different data structures and methods, and since our data has no censoring, recurrent events, etc, we chose the Harrell's Concordance index definition for the assessment of the accuracy of predictions, since it would be easier to understand with regards to our data and the use of random effects.

Let η_i be defined as in (16) and δ_i as in (13). The Harrell's Concordance index (C-index) [Schmid and Ziegler, 2016] is used as the measure of how well the model fits the data by the comparison of the pairs of individuals (i, j) from the Test data. It compares the risks (η_i, η_j) predicted on the individuals in the Test data based on the Train data and the actual event times corresponding to the individuals (T_i, T_j) in the Test data. The equation for Harrell's C-index is given below.

$$C = \frac{\sum_{i,j} I(T_i > T_j) \cdot I(\eta_j > \eta_i) \cdot \delta_j}{\sum_{i,j} I(T_i > T_j) \cdot \delta_j}, \tag{36}$$

where $\delta_j = 1$, since we have no censored events.

When the event time for the individual i happened after the event for individual j, that is, $I(T_i > T_j) = 1$, and the predicted risk for the individual j is higher than for the individual i, that is, $I(\eta_j > \eta_i) = 1$, then the prediction is accurate for that pair of (i, j), thus, the pair is concordant.

Therefore the Harrell's C-index corresponds to the probability $P(\eta_j > \eta_i | T_i > T_j)$ [Schmid and Ziegler, 2016]. The interpretation of the C-index is straightforward, as its value is the proportion of the concordant pairs over all pairs to compare, if $0 < C < 0.5$, then the prediction is not accurate; if $C = 0.5$, then the prediction is not informative; if $0.5 < C < 1$, then the model predicts the risks of events well.

4 Results

4.1 Preliminary analysis

In this section we provide the survival (Kaplan-Meier) curves for each of our fixed-effect variables, in order to provide visualization on the initial analysis and gain more intuition before the analysis of the estimated coefficients in the next subsection[13]. The corresponding R code is given in the Appendix C. The survival curves are informative in observing the survival probability as time (screening in theatres) increases and serve as the firm starting point for the comprehensive approach [Kim, 2021].

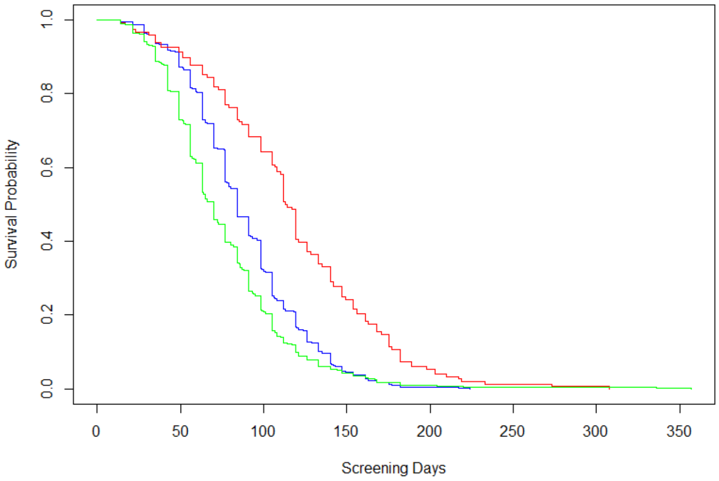

Figure 2: K-M curves for Rating. R (green); PG-13 (blue); PG & G (red)

For the **rating** of the movie (2), we observe that the survival curves for R (green) and PG-13 (blue) ranking movies are of the similar shape, compared to the PG & G (red) movies. This is connected to the similarity of movies with the age restrictions, compared to the movies with no age restriction stay. We see that the longest screening time are of movies with R rating, however, only for a few movies.

Within the first 60 days since the release, we notice that, regardless of the rating of the movie, more movies are likely to be withdrawn from the theatres with close to equal probability. After this period, the rating indicator for the movie seems to make the difference on the withdrawal from theatres.

According to the rate at which the survival probability decreases, we do observe that for R and PG-13 rating movies the rates are greater, however, more detailed analysis is required.

From the Kaplan-Meier curves for the **genres** (3), we observe that the Adventure/Western (red) and Musical/Concert/Performance (black) genres differ noticeably from the rest, and according to the plot, they have the longer screening days than the rest of the genres. We notice, however, that for the Musical/Concert/Performance, the amount of movies is a lot

[13] The software used for coding is R version 4.3.2 (2023-10-31 ucrt); RStudio environment

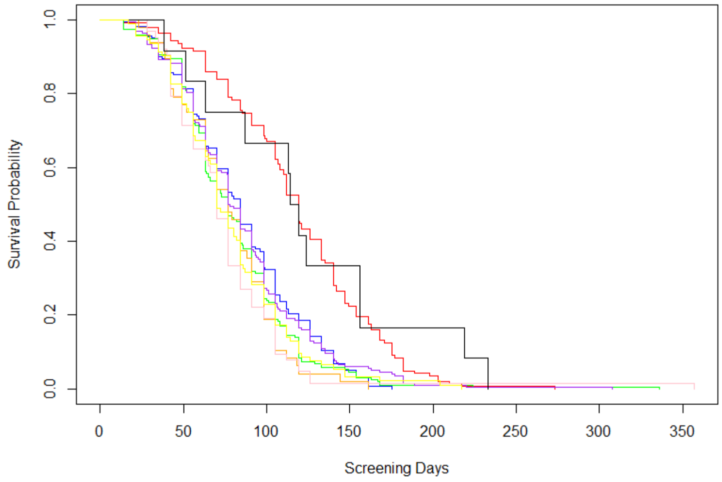

Figure 3: K-M curves for Genres. Adventure/Western (red); Musical/Concert/Performance (black); Action (blue); Drama (purple); Thriller/Crime (yellow); Comedy (green); Horror (pink); Documentary/Biography (orange)

less than for the rest (12 movies in total). This genre is popular for the specific audience, therefore due to the specific of this genre, it was also initially in question whether to include it into the research. It is interesting to notice that in many sources the genres Action (blue) and Adventure are grouped, however, we observe that Adventure is much more successful than Action, but the Action genre has higher survival probability as screening days increase, which is visible on the graph between 70 - 140 screening days, compared to the rest of genres. The Action genre "competes" with the Drama (purple).
Also, as the time increases, we observe the similarity between Thriller/Crime (yellow) and Comedy (green), being less successfully than the previous genres, but more successful than the rest.
The last genres Horror (pink) and Documentary/Biography (orange) tend to be withdrawn from theatres faster than the rest genres, however, the longest screening time within our time frame was the Horror movie "Get out!" scoring $255.4 million running in theatres for 357 days (excl.re-releases) [IMDB, 2024].

Regarding the variable **running time**, for the Categorization 1, (4) the survival curves are more difficult to analyze than for the Categorization 2 (5). For the 6 categories, although we see the differences between the behaviour of movie withdrawal at the different screening days range, the patterns seem to change. For example, the movies with longer run time, 116-126 min (pink) and 126-201 min (black), tend to have higher probability of survival than the rest of categories. However, as the amount of screening days increase (closer to 140-150 days), the probability of movie withdrawal is roughly the same for all categories. For the rest categories, 75 - 94 min (blue), 94-101 (red), 101-108 min (orange), 108-116 min (purple), the probabilities of survival are more difficult to compare.

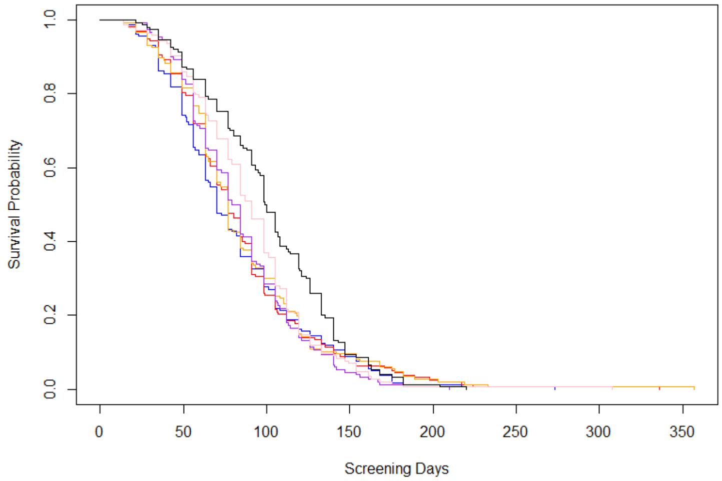

Figure 4: K-M curves for Runtime Categorization 1. 116-126 min (pink); 126-201 min (black); 75 - 94 min (blue), 94-101 (red), 101-108 min (orange), 108-116 min (purple)

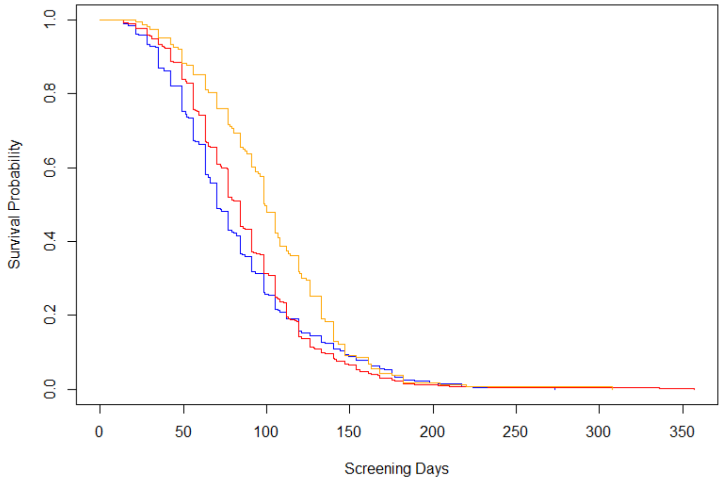

Figure 5: K-M curves for Runtime Categorization 2. Short movies (blue); Average length movies (red); Long movies (orange)

For the Categorization 2, we observe the clear differences between the withdrawal from theatres for the 3 categories constructed, especially in the first 30 - 110 days. The Short movies (blue), have the lowest probability of survival, followed by the Average length movies (red) and the longest survival probability for the Long movies (orange).

The **opening gross** (6) survival curves are more indicative and straightforward, compared to other variables. Despite the very low opening gross (blue) category, we observe that the

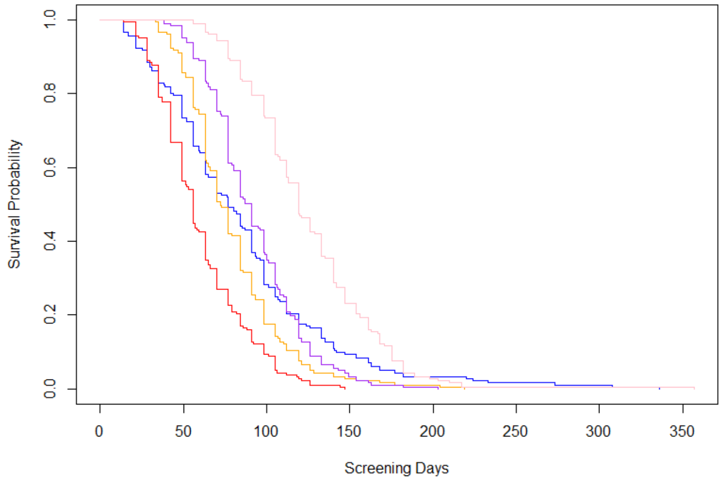

Figure 6: K-M curves for Opening Gross; Very low opening gross (blue); Low opening gross (red); Below average opening gross (orange); Above average opening gross (purple); High opening gross (pink)

survival probability of the movies is higher, as the opening gross increases. We highlight that the high opening gross (pink) is screened longer, followed by the above average opening gross (purple), then the below average opening gross (orange), and we have the low opening gross (red), with the lowest probability of survival as screening days increase.

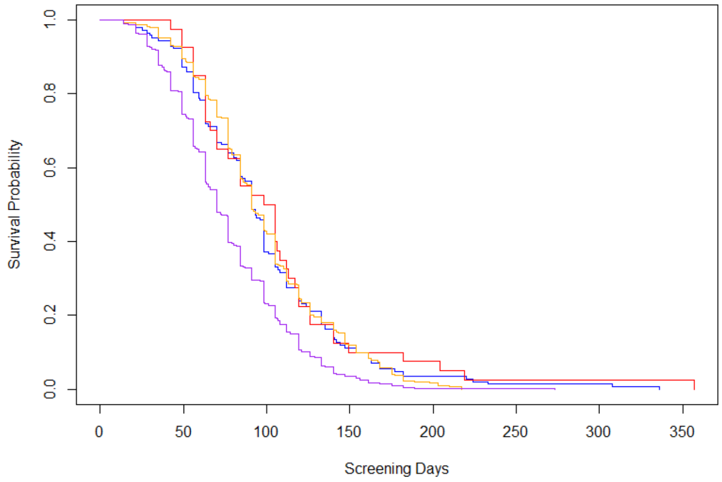

Figure 7: K-M curves for Open theatres. $<$ 25% (blue), 25 % to 90 % (red) , 90 % to 100 % (orange), 100 % (purple)

Finally, for the **amount of open theatres compared to the maximum theatres** (7), we notice that the movies with the 100% open theatres (purple) tend to have much lower survival probability than the rest of the categories. For more detailed examination we will require the regression results, since the survival plots are difficult to interpret for some of the covariates.

Thus, we can already construct some conclusions based on our analysis of the survival curves. However, in order to explore our hypotheses in a more comprehensive way, we proceed to the regression analysis.

4.2 Coxme

4.2.1 Regression results

Before reporting the fixed and random effect coefficients, we fit two versions of CoxME (Cox mixed effects model), one with Running time Category 1, the second with Running time Category 2. The AIC and BIC are reported below. The R code for this and further section is given in the Appendix B.

	Category 1	Category 2	Difference
AIC	9851.763	9846.291	$\Delta_{\text{AIC}} \approx 5$
BIC	10129.26	10109.88	$\Delta_{\text{BIC}} \approx 19$

Table 9: Comparison of Runtime Categories

From the table, we can see that the difference in BIC strongly suggests choosing Category 2 over Category 1. Therefore further in the paper we will proceed with the variable Running time Category 2. However, as described in the Data section, some of the information could be lost, therefore we also present the short discussion of the Category 1 fixed effects coefficients in the Appendix E.

The important observation upon fitting the CoxME in R is that what is called AIC and BIC in the regression output of the "coxme" is not the AIC and BIC of the model, calculated on the standard formulas. The values correspond to the tested models (inside the "coxme" algorithm). Thus, the reported AIC and BIC we discuss are the values in (9) which are calculated based on the standard formula and extracted using $AIC(coxmedatame)$ and $BIC(coxmedatame)$ functions in R.

On the next page is the CoxME regression results for the fixed effects coefficients (10). Based on the hazard ratios in exp(coef), having the comparison category for each covariate, we can better see the differences in the hazards. Regarding the interpretation of hazard rates for the Opening gross, we can clearly see that given the comparison category Very low opening gross, the hazards significantly decrease as the Opening gross values increase.

For the genres, given the comparison group Action, the hazard rates for genres are close to the value 1. In general, the analysis of the hazards correspond to the survival curves inspection. That suggest that the hazards across the genres are similar, however, we see the

	coef	exp(coef)	se(coef)	p
Opening Gross Low	-1.4811	0.2274	0.2025	2.6e-13
Opening Gross Below Average	-2.5243	0.08011	0.2186	0.0e+00
Opening Gross Above Average	-2.9523	0.0522	0.2282	0.0e+00
Opening Gross High	-3.8204	0.0220	0.2413	0.0e+00
genre Adventure/Western	-0.3042	0.7377	0.1515	4.5e-02
genre Comedy	-0.1967	0.8215	0.1198	1.0e-01
genre Documentary/Biography	0.1618	1.1756	0.1979	4.1e-01
genre Drama	-0.1931	0.8244	0.1287	1.3e-01
genre Horror	0.1279	1.1365	0.1736	4.6e-01
genre Musical/Concert/Performance	-0.2048	0.8149	0.3319	5.4e-01
genre Thriller/Crime	-0.0885	0.9154	0.1414	5.3e-01
% Open Theatres to Max 25-90%	2.0882	8.0700	0.2721	1.7e-14
% Open Theatres to Max 90-100%	3.0135	20.3586	0.2323	0.0e+00
% Open Theatres to Max 100%	3.3850	29.5187	0.2287	0.0e+00
rating PG&G	-0.6351	0.5299	0.1474	1.6e-05
rating R	0.2472	1.2805	0.0821	2.6e-03
running time Average	-0.1505	0.8603	0.0891	9.1e-02
running time Long	-0.4130	0.6617	0.1244	9.0e-04

Table 10: Fixed Effects Coefficients

lowest hazard for Adventure/Western and the highest for Documentary/Biography.

For the Open theatres, the hazard rate is huge for all categories, compared to the reference category of movies shown in less than 25% of theatres from their maximum on the opening day. Movies which were shown in their maximum amount of theatres on the release day have 29.5 times higher hazard. Movies which had 90% to 100% of screened theatres on release day have a 20.3 times higher hazard, followed by ones with 25% to 90% on release day with 8 times higher hazard.

Given the comparison category PG-13 rated movies, the PG&G rating is ($\frac{1}{0.5299} = 1.89$ times) more successful than the restrictive category PG-13. For the R rating, the movies have slightly shorter screening duration, compared to the PG-13.

Regarding the running time, a category long length of film has the lowest hazard, followed by the average length movies, but the differences of hazards might be insignificant due to the p-value of 9.1e-02 for the Average runtime. We will inspect this further.

For the Distributor, the summary for CoxME provides the estimated $\hat{\theta} = 0.6648$, which confirms that the Distributor can be treated as random effect, since the variation between movies under different Distributors is present. We note, that for the random effects coefficients, we do not have the reference/comparison group [Therneau, 2022]. Instead the comparison value is 0, so the comparison to 0 effect. The random effects coefficients are reported in the Appendix D.

Next, we compare the model fit between Mixed Effects Cox model and the standard Cox Proportional Hazard model (11). In the Cox PH we treat Distributors as fixed effects variable. We note, that when we implemented the Cox PH in R, we had to increase the number of iterations for the regression to estimate coefficients.

	CoxME	CoxPH	Difference
AIC	9846.291	9854.725	$\Delta_{\text{AIC}} \approx 8$
BIC	10109.88	10210.51	$\Delta_{\text{BIC}} \approx 101$

Table 11: Comparison with Standard Cox PH model

The difference in BIC suggest very strong preference of the mixed effects Cox regression model over the Cox Proportional Hazard model, hereby confirming our choice of Distributors as random effects over the standard approach.

4.2.2 Backward Elimination

Next we aim to identify which variables are significant for the model. We do this by performing the Backward elimination procedure, where the idea is to identify the subset of the variables which is most optimal for the model. In the general framework of the backward elimination, we first consider the model fitted with all of the X variables (10) and look at the variable with the highest p-value, usually above the threshold $p > 0.05$. The corresponding variable is removed from the model. The procedure is then repeated on the X - 1 variable model, and stops when no variables need to be removed [Kutner et al., 2005].

As indicated in [McQuire, 2024], the alternative measures to use are AIC and BIC values. Similarly, for the backward elimination, we drop the variables, one at a time, from the full model and compare the AIC and BIC values of the reduced models to the AIC and BIC of the full model. If the values of the AIC and BIC in the reduced model are higher, than in the full, the corresponding variable is significant to the model. If the values of AIC and BIC are lower, then the variable is removed from the model, and the model without this variable is fitted, repeating the AIC and BIC values comparison. The process stops when no variable removal contribute to the better model fit.

As described in [McQuire, 2024], using the AIC criterion for the backward elimination procedure, the first step of the algorithm[14] removes the variable for which the p-value is above the threshold and if the model without the corresponding variable induces the smallest AIC, then we remove this variable and proceed as described above.

The p-values above the threshold ($p > 0.05$) in the model (10) correspond to the Genre variable and the Running time. We first compare the full model with the reduced model without Genre (12).

	CoxME	CoxME excl. Genre	Difference
AIC	9846.291	9843.5	$\Delta_{\text{AIC}} \approx 3$
BIC	10109.88	10074.47	$\Delta_{\text{BIC}} \approx 35$

Table 12: Model comparison without genre

As we can see based on the thresholds for the difference in BIC values, the model without the genre is strongly preferred. Next, we compare the reduced model without the Runtime with the full model (13).

	CoxME	CoxME excl. Runtime	Difference
AIC	9846.291	9854.639	$\Delta_{\text{AIC}} \approx 8$
BIC	10109.88	10106.98	$\Delta_{\text{BIC}} \approx 3$

Table 13: Model comparison without runtime

Based on the AIC difference, the exclusion of the variable Runtime results in the worse model fit. However, we note that the difference in BIC suggests that removal of Runtime results in the somewhat better model, therefore the significance of the Runtime is partly cofirmed with AIC and BIC values. For the rest of the variables, we provide the table below.

	CoxME	excl. Open.gross	excl. Open.theaters	excl Rating
AIC	9846.291	10154.07	10069.54	9873.057
BIC	10109.88	10351.04	10277.58	10131.95

Table 14: Eliminating Opening gross, Percentage of Open theatres and Rating from the full model

[14] description of the algorithm in the book, which corresponds to the stepAIC function in R; We compute the AIC and BIC manually

Thus, we see, that Δ_{AIC} and Δ_{BIC} propose very strong support of the model with the inclusion of the remaining variables. Lastly, we remove the Distributor variable from the CoxPH model (since we cannot remove the random effect term in CoxME) to test the variable significance in the influence on the survival time. We report the following: AIC $=$ 10035.54; BIC $=$ 10122.09. Comparing to the AIC and BIC values of the full model we have strong support for the CoxME model with the Distributor, therefore the Distributor is a significant variable for our model.

For the next step of backward elimination, based on the difference in BIC values, the model without the Genre is chosen (12).

Thus, we now repeat the procedure, with the reduced model without the Genre, to test the significance of other variables. We start with the Runtime results summary, due to the partly supported effects in the previous step (15).

	CoxME without Genre	CoxME excl. Genre&Runtime	Difference
AIC	9843.5	9855.829	$\Delta_{\text{AIC}} \approx 12$
BIC	10074.47	10075.85	$\Delta_{\text{BIC}} \approx 1$

Table 15: Model comparison without Genre and Runtime

Based on the difference in AIC values, we have strong support of the model with the variable Runtime. Next, we assess the significance of the remaining variables.

	CoxME without Genre	excl. Open.gross	excl. Open.theaters	excl. Rating
AIC	9843.5	10181.3	10071.32	9906.799
BIC	10074.47	10342.3	10247.17	10140.57

Table 16: Eliminating Opening gross, Percentage of Open theatres and Rating from reduced model

The removal of every remaining variable resulted in the significantly worse fit. Further, testing the model without Distributor and genre, we have AIC $=$ 10034.84;
BIC $=$ 10087.73. Thus, the Distributor is significant to the model, together with the Opening gross, Percentage of Open theatres, Runtime and Rating, based on the second step of the backward elimination.

4.2.3 Cross-Validation

In the coxme package, the predict.coxme function does not yet support prediction on the test data [Therneau, 2024]. Therefore we installed the coxme helper function predict_coxme from the junkka/ehahelper package [Junkka, 2021]. Regarding the computation of the c-index, we searched for the functions which correspond to the Harrell's c-index [Gerds, 2023].

The results for the Harrell's C-indexes for different Test data subsets are given below.

	Test data 1	Test data 2	Test data 3	Test data 4	Test data 5
C-index	0.7627749	0.7792426	0.7193666	0.7693363	0.7647787

Table 17: 5-folds Cross-Validation

The Harrell's indexes suggest that CoxME model predicts the risks of the movies to be withdrawn well, which indicates the accuracy of the regression results.

Moreover, we decided to manually check the computation of the c-index in order to better understand the function cIndex in R and Harrell's c-index. The corresponding code is given in the Appendix B. We used 5-folds cross-validation, so the size of the Test data is 181, and manually we checked the outputs for the iteration when Test data is the 5th fold.

That is, the total number of pairs to compare is (using binomial coefficient): $\frac{181!}{2!\cdot 179!} = 16290$. We noticed that the c-index output gives the number of "comparable" pairs 15749, where the comparable means the pairs considered in the computation of the c-index. Such disparity between the number of pairs to compare and the actual comparable pairs followed from the exclusion of the pairs of the equal amount of screening days, when $T_i = T_j$. We checked this by identifying the movies with the same amount of screening days, calculating the amount of the respective combination of pairs for the same amount of screening days and summed over the amount of pairs (see Appendix B). The combination of pairs to compare with the same amount of screening days is 541, which corresponds to the difference between the number of pairs to compare and the actual comparable pairs with cIndex function. Thus, the pairs with the same amount of screening days are indeed not considered when computing the Harrell's index (36).

4.2.4 Validation of Normality Assumption

After extracting the random coefficients of Distributors from the estimated CoxME model, we assess the normal distribution of random effects coefficients. Below is the Q-Q plot, implemented in R, using qqPlot function, which compares estimated random effect coefficients to the standard normal distribution.

From the plot we can see that the distribution of random effects, in total 57 for every Distributor, is close to the normal distribution for most of the Distributors. However, the tails of the distribution deviate from the theoretical normal distribution line and the presence

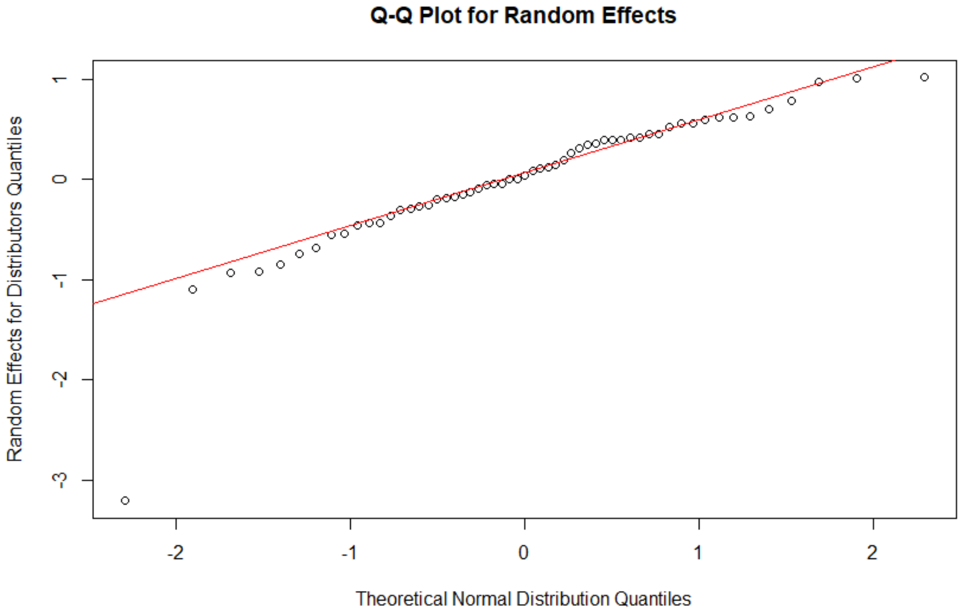

Figure 8: The Q-Q plot for the Distributor's Random Effects compared to the theoretical Normal distribution (red line)

of outliers suggest for further advanced tests in order to validate the normality assumption.

Our plot is also similar to [Sun et al., 2023] (the plots in the Supplementary material) in the presence of outliers, where the authors indicate that their plots are satisfactory, thus the normality assumption is not rejected.

5 Conclusion and discussion

5.1 Conclusion

Given the results in the previous subsection, we can now provide the comprehensive answer for our research question, namely: **"To what extent do movie production characteristics and release factors affect the success of movies?"**. By means of survival analysis, we explored our topic, specifically we used the extended Cox Regression Model with the inclusion of random effects to account for the clustered data structure, that is, by treating Distributors as the clusters. The Cox model with Mixed effects was shown to be preferred over the ordinary Cox regression model, by comparison of the AIC and BIC values. The assumption of normally distributed random effects for Distributors is assessed using q-q plot. The plot suggested that the distribution of random effects does not significantly deviate from the normal distribution. The Cox Mixed effects model is assessed with the cross-validation, using the Harrell's C-index as the measure for the accuracy of the model predictions on the Test data subset. The values of C-indexes suggest that the Cox Mixed effects model's predictions on the risks of the movies to be withdrawn from theatres are accurate. Below we answer our research question and the corresponding hypotheses on each of our variables, by using the results from the regression table (10), analysis of the survival curves and backward elimination for testing the significance of the variables.

The movie production characteristics such that **rating** and **distributor** affect the survival time of the movie in theatres. The inclusion of frailty for the Distributor is justified compared to the standard Cox proportional hazards model and the assessment of normality assumption. Regarding the rating, the movies with no age restriction (PG&G) have lower risk of withdrawal from theatres. This is indicative for some of the studios, to consider the exclusion of the mature content, if it suggests the higher movie success. For the theatres it is also informative in arranging the theatrical windows for movies with the less restrictive ratings.

Next, for the remaining production characteristic **running time**, the effect on the longevity of screening is partly supported, due to the more strict categorization of short/average/long movies, instead of analyzing the smaller intervals in film's length. The tendency of longer runtime being the most successful is observed with our results. For the **genres**, it being the indicative characteristics is not confirmed. Genres appear to have insignificant effect on the movie's screening duration, according to the regression results, survival curves plot and the backward elimination procedure.

Regarding the release factors, the **opening gross** and **the amount of open theatres compared to maximum theatres** appear to have a significant effect on the film's survival. The opening gross category is informative for the theatres, as for the studio analysis we would require the corresponding production budget. Thus, identifying in which opening gross category the film is, can influence the theatre's decision to extend the theatrical release. The higher opening gross does result in the lower risk of the movie withdrawal from theatres. Regarding the increase of the amount of theatres screening the movie as the release is ongoing, the studios can observe the success of the films, based on the wider release theatre amount, and adjust correspondingly the current and further movie projects or to arrange more theatre owners to screen the movie, as it appear to increase the film success. It was thus observed, that based on the regression results, the larger the increase in the theatres releasing the movie is, the longer the movie stays in theatres.

Based on the results of the above analyses, we can conclude that the film's characteristics such as rating, opening gross and the growth of the amount of theatres affect movies survival in theatres, and given that their exclusion from the model results in significantly worse fit, we have enough evidence to support **Hypotheses 1 & 4 & 5**. For the genres, we reject **Hypothesis 2**. Lastly, the runtime **Hypothesis 3** is partly supported (see Appendix E). As indicated at the beginning of the paper, our results can be used in further comparison to the mid-COVID and post-COVID industry periods, to analyze the change of film's trends.

5.2 Discussion

We note that variable categorization used in this project can be subjective, since the categorization can be done in different ways. Regarding the partly supported hypothesis for the runtime, we highlighted the importance of the appropriate categorization choice for the covariates, as discussed with two proposed categorizations. In our work we attempted to refer to the accessible and easily interpreted sources for the categorization/grouping for some variables, however, the resulting categorization is not optimal, which is a note for further improvement.

The approach for examining genres could have been improved. We believe that it is possible to get a more comprehensive inference for the genres, i.e. considering multiple genres of the film instead of only the predominant. Another potential improvement of the objectivity of the categorization is to reference film theory in-depth. We note that in the referenced research papers the effect of genre on the survival was observed, in opposite to our paper.

The rating variable is based on MPAA classifications [MPA, 2020], which can be considered objective. However, some criticism can be applied regarding the subjectivity of the MPAA ratings definitions. For instance the terms such as "adult themes" and "disturbing imagery" are vague and the meaning of these terms might differ per country. For example, according to [IMDB, 2024], the domestic (US) rating for "Harry Potter and the Deathly Hallows: Part 2" is PG-13, while for the Ukrainian theatrical release the rating 16+ is assigned [Multiplex, 2024].

The categorization of the theatres variable aims to explore the effect of the increase of screening theatres on a film's survival. However, the obtained data did not have specific enough information on daily/weekly increase, thus, we took the number of opening theatres to the maximum number of theatres (in percentage of the opening theatres compared to maximum). The tracking of the growth of the amount of theatres could have improved the inference for this variable.

Regarding the cross-validation, the exclusion of the pairs with the same amount of screening days, we suggest that the there are more suitable measures of the model prediction than the Harrell's c-index. Based on the study which proposes several model predictive measures [Dai and Breheny, 2019], in addition to the Harrell's c-indexes, the authors suggest the mean squared error of the coefficients (MSE); Brier score and the Kullback-Leibler score. Thus, in the further research, one of these measures can be tested for suitability for coxme prediction. Moreover, regarding the R function coxme_predict, since the function is from the helper package (not official), ones the official function releases, its usage is more appropriate.

Concerning the model suitability, we note, that when comparing the models CoxME and CoxPH we noticed that the degrees of freedom for the coxme model is 54.82. For the CoxPH model, the degrees of freedom are 74 (18 + 56 excl. the comparison categories also for Distributors). We note, that according to the documentation of Therneau coxme package, the coxme degrees of freedom are the effective degrees of freedom, which "lie somewhere between" the total number of coefficients (75 (fixed effects excl comparison categories + 57 distributors)) and the degrees of freedom for the integrated partial likelihood (18 + 1 for the random effects intercept (variance)) [Therneau, 2022]. So the effective degrees of freedom differs from the degrees of freedom, and due to the effective degrees of freedom, the Coxme was preferred over the ordinary Coxph by comparing AIC and BIC. [Therneau and Grambsch, 2000] indicates that the estimated variance $\hat{\theta}$ is used in the calculation of the degrees of freedom. We indicate that further explanations and discussion for this procedure is needed, as this might be the important discussion, which goes beyond our scope. To summarize, as the algorithm for estimating coefficients and variance goes as the outer loop with maximizing the PPL and IPL, the effective degrees of freedom lie between their degrees of freedom.

We note that for the general description of the Cox mixed effects estimation we took as the starting point the documentation for coxme [Therneau, 2022], but in the main body of the work we explain mathematical concepts referencing the book by Therneau [Therneau and Grambsch, 2000] and the article by Ripatti [Ripatti and Palmgren, 2000], in order not to confuse the accuracy of the definitions of the likelihood functions. Since the documentation is not the official reference literature (no doi nor ISBN), the definitions there complicated the initial understanding, possibly due to the optimizations for the construction of the coxme function, stated/mentioned in the documentation, but without the in-depth description.

Throughout this project we tried to expand the applications of survival analysis in the film industry. As the data driven approaches are integrated in the entertainment industry, this leads to the further possibilities to analyze the film's data by using other statistical methodologies or improving the existing researches with the new variables and data.

References

[MPA, 2020] (2020). *Classification and Rating Rules.* Motion Picture Association Inc. and National Association of Theatre Owners Inc. pp 5-7. https://www.filmratings.com/RatingsGuide.

[Altman, 1999] Altman, R. (1999). *Film/Genre.* British Film Institute. pp. 14-17. isbn:13: 9780851707181.

[Austin, 2017] Austin, P. C. (2017). A tutorial on multilevel survival analysis: Methods, models and applications. *International Statistical Review*, 85(2). https://doi.org/10.1111/insr.12214.

[Berchtold, 2010] Berchtold, A. (2010). Sequence analysis and transition models. *Encyclopedia of Animal Behavior.* https://doi.org/10.1016/B978-0-08-045337-8.00233-3.

[Burnham and Anderson, 2004] Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference. *Sociological Methods & Research.* https://doi.org/10.1177/004912410426864.

[Casella and Berger, 2002] Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Cengage Learning.

[Chisholm and Norman, 2006] Chisholm, D. C. and Norman, G. (2006). When to exit a product: Evidence from the u.s. motion-picture exhibition market. *American Economic Review*, 96(2). https://doi.org/10.1257/000282806777212440.

[Dai and Breheny, 2019] Dai, B. and Breheny, P. (2019). Cross validation approaches for penalized cox regression. *Department of Biostatistics, University of Iowa.* https://doi.org/10.48550/arXiv.1905.10432.

[Dobson and Barnett, 2018] Dobson, A. J. and Barnett, A. G. (2018). *An Introduction to Generalized Linear Models.* CRC Press, 4 edition. pp. 223-240; 272-273; 287. isbn 9781138741683.

[Dodge, 2008] Dodge, Y. (2008). Q-q plot (quantile to quantile plot). The Concise Encyclopedia of Statistics. https://doi.org/10.1007/978-0-387-32833-1_331.

[Fitzmaurice et al., 2011] Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis.* John Wiley & Sons, 2 edition. pp 410-411.

[Gerds, 2023] Gerds, T. A. (2023). cindex: Concordance index, intsurv package. https://www.rdocumentation.org/packages/intsurv/versions/0.2.2/topics/cIndex.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. 2 edition. pp. 241-247 ISBN 978-0-387-84858-7.

[Huang, 2022] Huang, P. (2022). Using stepwise method to find the most influencing feature to the cell nuclei of a breast mass. *2022 International Conference on Applied Mathematics, Modeling Simulation and Automatic Control (AMMSAC 2022)*, 16. https://doi.org/10.54097/hset.v16i.2632.

[IMDB, 2024] IMDB (2024). Box office mojo. Information courtesy of IMDb https://www.boxofficemojo.com/year/2018/?grossesOption=totalGrosses&sort=releaseDate&ref_=bo_yld__resort#table.

[Junkka, 2021] Junkka, J. (2021). ehahelper: R helper package for survival analysis. https://rdrr.io/github/junkka/ehahelper/.

[Kim, 2021] Kim, Ahyun, e. a. (2021). Exploring the key success factors of films: A survival analysis approach. *Service Business*. https://doi.org/10.1007/s11628-021-00460-x.

[Kleinbaum and Klein, 2005] Kleinbaum, D. G. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. Springer, 2nd edition.

[Kutner et al., 2005] Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5 edition. ISBN 0-07-238688-6, p. 368.

[Legoux et al., 2015] Legoux, R., Larocque, D., Laporte, S., Belmati, S., and Boquet, T. (2015). The effect of critical reviews on exhibitors' decisions: Do reviews affect the survival of a movie on screen? *International Journal of Research in Marketing*. https://doi.org/10.1016/j.ijresmar.2015.07.003.

[Los Angeles Times, 2023] Los Angeles Times (2023). Writers strike 2023: What to know about the wga's fight and how it could upend hollywood productions. https://www.latimes.com/entertainment-arts/business/story/2023-05-01/writers-strike-what-to-know-wga-guild-hollywood-productions.

[MacKay, 2003] MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. Chapter 27, pp. 341-342 ISBN-13: 9780521642989.

[McQuire, 2024] McQuire, P., K. A. (2024). *R programming for Actuarial Science*. John Wiley & Sons Ltd. ISBN 9781119754985, pp 123-128.

[Multiplex, 2024] Multiplex (2024). Theatrical releases. Information of the old releases https://multiplex.ua/ru/movie/353961.

[Nash, 2023] Nash, N. I. S. (2023). Opus data movie data extract. https://www.opusdata.com/.

[NoFilmSchool, 2022] NoFilmSchool (2022). We're releasing roughly 30% fewer movies in theaters. https://nofilmschool.com/releasing-fewer-movies.

[Numbers, 2023] Numbers, T. (2023). Film industry data website operated by nash. https://www.the-numbers.com/.

[OpusData, 2023] OpusData, N. I. S. (2023). Opus data documentation. www.opusdata.com/documentation/index.php/Movie_theatrical_release.

[Pokorny et al., 2019] Pokorny, M., Miskell, P., and Sedgwick, J. (2019). Managing uncertainty in creative industries: Film sequels and hollywood's profitability 1988–2015. *Competition & Change*. pp. 24-27 https://doi.org/10.1177/1024529418797302.

[Ripatti and Palmgren, 2000] Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*. https://doi.org/10.1111/j.0006-341x.2000.01016.x.

[Rosser, 2019] Rosser, M. (2019). Does a long running time help or hurt a film's box office performance? *Screen Daily*. https://www.screendaily.com/features/does-a-long-running-time-help-or-hurt-a-films-box-office-performance/5144271.article.

[Rteam, 2019] Rteam (2019). *qqPlot Function Documentation*. RDocumentation. https://www.rdocumentation.org/packages/EnvStats/versions/2.3.1/topics/qqPlot.

[Schmid and Ziegler, 2016] Schmid, M. N. W. and Ziegler, A. (2016). On the use of harrell's c for clinical risk prediction via random survival forests. https://doi.org/10.48550/arXiv.1507.03092.

[Sun et al., 2023] Sun, Y., Zhou, Q., and Gilbert, P. (2023). Analysis of the cox model with longitudinal covariates with measurement errors and partly interval censored failure times, with application to an aids clinical trial. *Statistical Biosciences*. Analysis in the Supplementary Information. https://doi.org/10.1007/s12561-023-09372-y.

[Therneau, 2024] Therneau (2024). predict.coxme for coxme package. not yet supported on 28/01/2024 https://rdrr.io/cran/coxme/man/predict.coxme.html.

[Therneau, 2022] Therneau, T. M. (2022). *Package 'coxme': Mixed Effects Cox Models; Coxme and the Laplace Approximation*. Mayo Clinic. https://cran.r-project.org/web/packages/coxme/index.html.

[Therneau and Grambsch, 2000] Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer Science Business Media. ISBN 0-387-98784-3.

[Zhang, 2006] Zhang, X. (2006). Generalized estimating equations for clustered survival data. A dissertation submitted to the graduate faculty. Program of Study Committee: Kenneth Koehler, Co-major Professor Terry Therneau, Co-major Professor Mervyn Marasinghe Max Morris Richard Evans. https://doi.org/10.31274/rtd-180813-16530.

A Appendix

B R code for fitting Cox mixed effects and Cox PH

Additionally, below we provide code for variable categorization, backward elimination and Q-Q plot and cross-validation

```r
install.packages("survival")
install.packages("readxl")
install.packages("devtools")
devtools::install_github('junkka/ehahelper')
install.packages("coxme")
install.packages("intsurv")

library(ehahelper)
library(coxme)
library(survival)
library(readxl)
library(intsurv)

datame <- read_excel("C:/Users/Maria/Desktop/thesis/DATA/Final/final.xlsx")

#We note that we split code over multiple lines in order for better
#readability in the thesis

#G and PG ratings
datame$rating <- as.character(datame$rating)
datame$rating <- ifelse(datame$rating %in% c("PG", "G"), "PG&G", datame$rating)
table(datame$rating)

#Grouping genres
datame$genre <- as.character(datame$genre)
datame$genre <- ifelse(datame$genre %in% c("Adventure", "Western"),
                       "Adventure/Western",
                       ifelse(datame$genre %in% c("Documentary", "Biography"),
                              "Documentary/Biography",
                              ifelse(datame$genre %in% c("Comedy",
                                                         "Romantic Comedy",
                                                         "Black Comedy"),
                                     "Comedy",
                                     ifelse(datame$genre %in%
                                              c("Musical", "Concert/Performance"),
                                            "Musical/Concert/Performance",
                                            ifelse(datame$genre %in%
                                                     c("Thriller/Suspense",
                                                       "Crime"),
                                                   "Thriller/Crime",
                                                   datame$genre))
                              )
                       )
)
table(datame$genre)

#categorizing Running time (Categorization 2)
```

```
datame$running_time <- as.numeric(datame$running_time)
datame$running_time <- cut(datame$running_time, breaks =
                             c(75, 99, 124, 201),
                           labels = c("Short Runtime", "Average Runtime",
                                      "Long Runtime"),
                           include.lowest = TRUE)

table(datame$running_time)

quantiles <- quantile(datame$Opening_Gross, probs = c(0, 0.2, 0.4, 0.6, 0.8, 1))
print(quantiles)

#categorizing Opening Gross
datame$Opening_Gross <- as.numeric(datame$Opening_Gross)
datame$Opening_Gross <- cut(datame$Opening_Gross, breaks =
                              quantile(datame$Opening_Gross,
                                       probs = c(0, 0.2, 0.4, 0.6, 0.8, 1)),
                            labels = c("Very Low Opening Gross",
                                       "Low Opening Gross",
                                       "Below Average Opening Gross",
                                       "Above Average Opening Gross",
                                       "High Opening Gross"),
                            include.lowest = TRUE)
table(datame$Opening_Gross)

# categorizing percentage of Open Theatres
datame$Percentage_of_Open_theatres_to_Max <-
  as.numeric(datame$Percentage_of_Open_theatres_to_Max)

datame$Percentage_of_Open_theatres_to_Max <-
  cut(datame$Percentage_of_Open_theatres_to_Max,
      breaks = c(0, 25, 90, 100, 100.0001),
      labels = c("< 25%", "25% to 90%",
                 "90% to 100%", "100%"),
      right = FALSE)
table(datame$Percentage_of_Open_theatres_to_Max)

datame$Distributor <- as.factor(datame$Distributor)
datame$Distributor

datame$Amount_of_screening_days <- as.numeric(datame$Amount_of_screening_days)

#Cox model with mixed effects.
#since the event happened for all observations, censoring = 1
coxmedatame <- coxme(Surv(time = Amount_of_screening_days,
                          event = rep(1, nrow(datame)))
                     ~ Opening_Gross + genre +
                       Percentage_of_Open_theatres_to_Max +
                       rating + running_time + (1|Distributor), data = datame)
coxmedatame
AIC(coxmedatame)
BIC(coxmedatame)
logLik(coxmedatame)

#random effects coefficients from coxme for each Distributor
```

```
redistributors <- ranef(coxmedatame)
redistributors

#unsuccessful coxph
coxphunsuc <- coxph(Surv(time = Amount_of_screening_days,
                         event = rep(1, nrow(datame)))
                    ~ Opening_Gross + genre + Percentage_of_Open_theatres_to_Max
                    + rating + running_time + Distributor, data = datame)

#coxph increasing itearations
coxphmodel <- coxph(formula = Surv(time = Amount_of_screening_days,
                                   event = rep(1, nrow(datame))) ~ Opening_Gross
                    + genre + Percentage_of_Open_theatres_to_Max +
                      rating + running_time + Distributor, data = datame,
                    control = coxph.control(iter.max = 50))
AIC(coxphmodel)
BIC(coxphmodel)
logLik(coxphmodel)
coxphmodel

#backward elimination; step 1
coxmedatamegenre <- coxme(Surv(time = Amount_of_screening_days,
                               event = rep(1, nrow(datame)))
                          ~ Opening_Gross + Percentage_of_Open_theatres_to_Max +
                            rating + running_time + (1|Distributor),
                          data = datame)
AIC(coxmedatamegenre)
BIC(coxmedatamegenre)


coxmedatameruntime <- coxme(Surv(time = Amount_of_screening_days,
                          event = rep(1, nrow(datame)))
                     ~ Opening_Gross + genre +
                       Percentage_of_Open_theatres_to_Max +
                       rating + (1|Distributor), data = datame)
AIC(coxmedatameruntime)
BIC(coxmedatameruntime)


coxmedatameopeninggross <- coxme(Surv(time = Amount_of_screening_days,
                                      event = rep(1, nrow(datame)))
                                 ~  genre + Percentage_of_Open_theatres_to_Max +
                                   rating + running_time + (1|Distributor),
                                 data = datame)
AIC(coxmedatameopeninggross)
BIC(coxmedatameopeninggross)

coxmedatameopenth <- coxme(Surv(time = Amount_of_screening_days,
                                event = rep(1, nrow(datame)))
                           ~ Opening_Gross + genre  +
                             rating + running_time + (1|Distributor),
                           data = datame)
AIC(coxmedatameopenth)
BIC(coxmedatameopenth)

```

```
coxmedatamerating <- coxme(Surv(time = Amount_of_screening_days,
                                event = rep(1, nrow(datame)))
                           ~ Opening_Gross + genre +
                             Percentage_of_Open_theatres_to_Max
                           + running_time + (1|Distributor), data = datame)
AIC(coxmedatamerating)
BIC(coxmedatamerating)


#without distributor coxph
coxphmodeldistrib <- coxph(formula = Surv(time = Amount_of_screening_days,
                                          event = rep(1, nrow(datame))) ~
                             Opening_Gross
                           + genre + Percentage_of_Open_theatres_to_Max +
                             rating + running_time, data = datame,
                           control = coxph.control(iter.max = 50))
AIC(coxphmodeldistrib)
BIC(coxphmodeldistrib)


#backward elimination step 2
coxmedatameruntimegenre <- coxme(Surv(time = Amount_of_screening_days,
                                      event = rep(1, nrow(datame)))
                                 ~ Opening_Gross +
                                   Percentage_of_Open_theatres_to_Max +
                                   rating + (1|Distributor), data = datame)
AIC(coxmedatameruntimegenre)
BIC(coxmedatameruntimegenre)


coxmedatamegenreopengross <- coxme(Surv(time = Amount_of_screening_days,
                                      event = rep(1, nrow(datame)))
                                 ~ Percentage_of_Open_theatres_to_Max +
                                   rating + running_time + (1|Distributor),
                                 data = datame)

AIC(coxmedatamegenreopengross)
BIC(coxmedatamegenreopengross)


coxmedatamegenreopenth <- coxme(Surv(time = Amount_of_screening_days,
                          event = rep(1, nrow(datame)))
                     ~ Opening_Gross  +
                       rating + running_time + (1|Distributor), data = datame)
AIC(coxmedatamegenreopenth)
BIC(coxmedatamegenreopenth)

coxmedatamegenrerating <- coxme(Surv(time = Amount_of_screening_days,
                          event = rep(1, nrow(datame)))
                     ~ Opening_Gross +
                       Percentage_of_Open_theatres_to_Max
                     + running_time + (1|Distributor), data = datame)
AIC(coxmedatamegenrerating)
BIC(coxmedatamegenrerating)

```

```
coxphmodelgenredistrib <- coxph(formula = Surv(time = Amount_of_screening_days,
                                          event = rep(1, nrow(dame))) ~
                             Opening_Gross
                           + Percentage_of_Open_theatres_to_Max +
                             rating + running_time, data = dame,
                           control = coxph.control(iter.max = 50))
AIC(coxphmodelgenredistrib)
BIC(coxphmodelgenredistrib)

#Q-Q plot; assessing the normality assumption
typeof(redistributors)
revector <- unlist(redistributors)

library(EnvStats)
qqPlot(revector, main = "Q-Q Plot for Random Effects",
       xlab = "Theoretical Normal Distribution Quantiles", ylab =
         "Random Effects for Distributors Quantiles")
qqline(revector, col = "red")


#cross-validation
#prepare fold number to assign for movies in data (as 1,2,3,4,5,1,2,3,4,...)
folds <- rep(1:5, length.out = nrow(dame))

cindexes <- numeric(5)

for (i in 1:5) {
  traindata <- dame[folds != i, ]
  testdata <- dame[folds == i, ]

  coxmetrain <- coxme(Surv(time = Amount_of_screening_days,
                           event = rep(1, nrow(traindata)))
                      ~ Opening_Gross + genre +
                        Percentage_of_Open_theatres_to_Max +
                        rating + running_time + (1|Distributor),
                      data = traindata)

  predictme <- predict_coxme(coxmetrain, newdata = testdata)

  cindexes[i] <- cIndex(testdata$Amount_of_screening_days,
                        rep(1, nrow(testdata)), predictme)

}
cindexes

#manually  checking number of concordant pairs
folds1 <- rep(1:5, length.out = nrow(dame))
traindata1 <- dame[folds1 != 5, ]
testdata1 <- dame[folds1 == 5, ]

coxmetrain1 <- coxme(Surv(time = Amount_of_screening_days,
                          event = rep(1, nrow(traindata1)))
                     ~ Opening_Gross + genre +
                       Percentage_of_Open_theatres_to_Max +
                       rating + running_time + (1|Distributor),
                     data = traindata1)

predictme1 <- predict_coxme(coxmetrain1, newdata = testdata1)

cindex1 <- cIndex(testdata1$Amount_of_screening_days,
                  rep(1, nrow(testdata1)), predictme1)
cindex1

#check for pairs with the same # of screening days
sameamountcheck <- table(testdata1$Amount_of_screening_days)
sameamountcheck

#note, if 1 movie for the amount of screen days, then pairssameamount=0
pairssameamount <- (sameamountcheck*(sameamountcheck - 1))/2
pairssameamount

totalpairssameamount <- sum(pairssameamount)
totalpairssameamount

#running time (Categorization 1)
#datame$running_time <- as.numeric(datame$running_time)
#datame$running_time <- cut(datame$running_time, breaks =
#c(75, 94, 101, 108, 116, 126, 201),
#labels = c("Group 1", "Group 2",
#           "Group 3", "Group 4",
#           "Group 5", "Group 6"),
#include.lowest = TRUE)
```

C R code for survival curves

```
#note: we also group the variables as in coxme before plotting survival curves
#note 2: for running time for different categories we re-categorize
#before plotting Category 2

#survival curves for ratings
kmrating <- survfit(Surv(time = Amount_of_screening_days,
                          event = rep(1, nrow(datame))) ~ rating, data = datame)
plot(kmrating, col = c("blue", "red", "green"), xlab = "Screening Days",
     ylab = "Survival Probability")
table(datame$rating)

#survival curves for genres
kmgenre <- survfit(Surv(time = Amount_of_screening_days, event =
                           rep(1, nrow(datame))) ~ genre, data = datame)
plot(kmgenre, col = c("blue", "red", "green", "orange", "purple",
                       "pink", "black", "yellow", "brown"),  xlab =
       "Screening Days", ylab = "Survival Probability")
table(datame$genre)

#survival curves for Running time (Categorization 1)
kmruntime <- survfit(Surv(time = Amount_of_screening_days,
                           event = rep(1, nrow(datame))) ~ running_time,
                     data = datame)
plot(kmruntime, col = c("blue", "red", "orange", "purple", "pink", "black"),
     xlab = "Screening Days", ylab = "Survival Probability")
levels(datame$running_time)

#survival curves Running time (Categorization 2)
kmruntime <- survfit(Surv(time = Amount_of_screening_days, event =
                             rep(1, nrow(datame))) ~ running_time, data = datame)
plot(kmruntime, col = c("blue", "red", "orange"),  xlab = "Screening Days",
     ylab = "Survival Probability")
levels(datame$running_time)

#survival curves for Opening Gross
kmopeninggross <- survfit(Surv(time = Amount_of_screening_days, event =
                                  rep(1, nrow(datame))) ~ Opening_Gross,
                          data = datame)
plot(kmopeninggross, col = c("blue", "red", "orange", "purple", "pink"),
     xlab = "Screening Days", ylab = "Survival Probability")
levels(datame$Opening_Gross)

#survival curves for Open theatres (compared to maximum theatres)
kmopenthtomax <- survfit(Surv(time = Amount_of_screening_days, event =
                            rep(1, nrow(datame))) ~
                           Percentage_of_Open_theatres_to_Max,
                     data = datame)
plot(kmopenthtomax, col = c("blue", "red", "orange", "purple"),
     xlab = "Screening Days", ylab = "Survival Probability")
levels(datame$Percentage_of_Open_theatres_to_Max)
```

D Distributors random effects

Below is the table of the estimated random effects for each of the 57 Distributors.

Table 18: Random Effects Coefficients for Distributors

Distributor	Coefficients
Affirm Films	-0.1482
Amazon Studios	0.6119
Annapurna Pictures	0.1169
ArtAffects Entertainment	-0.9246
Atlas Distribution Company	-0.1973
Aviron Pictures	-0.0953
BH Tilt	0.5576
Bleecker Street Media	0.0034
Briarcliff Entertainment	0.3949
Broad Green Pictures	-0.4541
CBS Films	0.7774
Cinelou Films	0.0794
Clarius Entertainment	0.4146
Dimension Films	-0.2887
Electric Entertainment	-0.0491
Entertainment Studios Motion Pictures	-1.0993
EuropaCorp	-0.0518
FilmDistrict	0.3593
Freestyle Releasing	0.0047
Fun Academy Motion Pictures	-3.2061
GKIDS	0.4503
Global Road Entertainment	-0.7394
IFC Films	-0.8476
LD Entertainment	0.4522
Metro-Goldwyn-Mayer (MGM)	-0.2709
Millennium Entertainment	0.3466
Mirror/LD Entertainment	-0.1846
Neon	0.5255

Ocean Avenue Entertainment	0.3907
Pantelion Films	0.9770
Pinnacle Peak	-0.3653
Purdie Distribution	-0.5521
Pure Flix Entertainment	-0.9312
Quality Flix	-0.0400
RADiUS-TWC	0.1028
RCR Distribution	0.5556
Roadside Attractions	-0.3103
Screen Gems	0.5982
Sony Pictures Classics	-0.4405
Studio 8	0.1475
The Samuel Goldwyn Company	0.3930
TriStar Pictures	0.1882
Viva Pictures	0.6317
A24	-0.2609
Focus Features	0.2578
Fox Searchlight	-0.6824
Lions Gate Films	0.6948
Open Road Films (II)	0.3092
Paramount Pictures	1.0129
Relativity Media	-0.4369
Sony Pictures Entertainment (SPE)	0.0347
STX Entertainment	1.0050
The Weinstein Company	-0.5448
Twentieth Century Fox	-0.1329
Universal Pictures	0.6182
Walt Disney Studios Motion Pictures	-0.1792
Warner Bros.	0.4214

E Fixed effects for Runtime Category 1

Even though we have chosen Runtime Category 2 over Runtime Category 1, here we report the fixed effects coefficients for the Running time Category 1.

	coef	exp(coef)	se(coef)	p
running time Group 2	-0.1428	0.8668	0.1260	2.6e-01
running time Group 3	-0.1018	0.9031	0.1302	4.3e-01
running time Group 4	-0.1818	0.8337	0.1338	1.7e-01
running time Group 5	-0.1833	0.8324	0.1400	1.9e-01
running time Group 6	-0.4652	0.6279	0.1471	1.6e-03

Table 19: Regression results for Running time Category 1

Compared to Group 1 (the smallest runtime), we observe that all other groups have the longer screening time. We see that there is indeed the tendency for the decrease in hazard rate as the running time increases, but the differences in hazard rates are small. That is, only for the longest run time, Group 6, we observe more significant decrease in hazard rate. Moreover, Group 3 hazard is higher than for Group 2, despite Group 3 having longer running time, therefore this confirms, that the choice of Categorization 2 is preferred.

With the Categorization 2, it was observed that, according to AIC value, the model with the Runtime is preferred. With the Categorization 1, we observe that the model without the Runtime and Genre is preferred, based on the BIC values, as shown below.

	CoxME	CoxME without Genre	CoxME without Genre&Runtime
AIC	9851.763	9848.453	9855.829
BIC	10129.26	10093.44	10075.85

Table 20: Model comparison for Categorization 1

These results underline the importance of accurate categorization of variables.