What are the main factors affecting movie profitability?

KARL WALLSTRÖM

MARKUS WAHLGREN
What are the main factors affecting movie profitability?

KARL WALLSTRÖM
MARKUS WAHLGREN

Degree Projects in Applied Mathematics and Industrial Economics
Degree Programme in Industrial Engineering and Management
KTH Royal Institute of Technology year 2018
Supervisors at KTH: Daniel Berglund, Julia Liljegren
Examiner at KTH: Henrik Hult
Abstract

This thesis consists of two parts which both focus on the profitability of movies with a theatrical release. The first part has examined if it is possible to create a model for predicting and understanding the profitability of movies. Factors that influence the profitability have also been identified. This has been achieved through a regression analysis with data on movies from 2015-2017. A final multiple regression model has been created which explains 41% of the variability in the data. Significant influencing factors with at least a level of 5% significance have been identified such as the genres Horror and Musical, the MPAA-ratings PG and PG-13, the creative types Historical Fiction and Dramatization, Rotten Tomatoes Score and Sequel.

The second part is an analysis on movie theaters as a distribution channel with respect to profitability for distribution companies using Porter’s five forces. The five forces analysis suggests that movie theaters are a good distribution channel for studios which can provide mainstream movies while studios should consider releasing their low profile movies directly to streaming services since it might be more profitable.
Sammanfattning

Acknowledgements

We would like to thank OpusData for supplying us with the necessary data making this thesis possible.
Contents

1 Introduction .................................................. 9
   1.1 Background ................................................. 9
   1.2 Purpose and motivation ................................ 10
   1.3 Disposition ............................................... 10
   1.4 Scope ...................................................... 10
   1.5 Problem Statement ...................................... 11

2 Mathematical Theory ........................................ 11
   2.1 Multiple linear regression ............................... 11
   2.2 Ordinary Least Squares ................................ 12
   2.3 Assumptions ................................................. 12
   2.4 Dealing with violation of assumptions ................. 13
       2.4.1 Residuals vs fitted ............................... 13
       2.4.2 Residuals vs regressors plot .................... 13
       2.4.3 Q-Q plot ........................................... 13
       2.4.4 Histogram of residuals ........................... 13
       2.4.5 Scale-location plot ............................... 13
       2.4.6 Box-Cox Transform ............................... 14
   2.5 Multicollinearity ......................................... 14
       2.5.1 VIF ................................................... 14
   2.6 Detecting leverage and influential observations .... 14
       2.6.1 Cook's distance ................................... 15
   2.7 Hypothesis testing ....................................... 15
       2.7.1 F-Statistic .......................................... 15
       2.7.2 t-test ................................................. 16
   2.8 Variable selection and model selection ............... 16
       2.8.1 All possible regression ........................... 16
       2.8.2 Cross-validation ................................... 17
       2.8.3 R² and adjusted R² ................................ 17
       2.8.4 Mallows's C_p statistic .......................... 17
       2.8.5 AIC .................................................. 17
   2.9 Dummy variables .......................................... 18

3 Method ......................................................... 18
   3.1 Data collection ........................................... 18
   3.2 Software .................................................. 18
   3.3 Dataset from Opus Data .................................. 18
   3.4 Preprocessing of Opus Dataset .......................... 19
   3.5 Creation of new variables ............................... 20
   3.6 Initial transformation of variables ..................... 20
   3.7 Choice of response variable ............................. 21
   3.8 Choice of regressors ..................................... 21
   3.9 Initial model .............................................. 21
       3.9.1 Residuals plotted against the fitted values .... 24
3.9.2 The residuals plotted against the regressor variables 25
3.9.3 Normal Q-Q plot 26
3.9.4 Histogram of residuals 26
3.9.5 Scale-Location plot 27
3.9.6 Residual vs leverage plot 28
3.9.7 Multicollinearity 28
3.10 Transformation of the model 29
3.11 Reduction of the model 29
3.11.1 Models generated by the first approach 29
3.11.2 Models generated by the second approach 30
3.11.3 Cross validation results 30
3.11.4 Choice, verification and transformation of final model 30

4 Result 30
4.1 Final model 30
4.1.1 Residuals plotted against the fitted values 33
4.1.2 Normal Q-Q plot 34
4.1.3 Histogram of residuals 34
4.1.4 Scale-location plot 35
4.1.5 Residual vs leverage plot 36
4.1.6 Multicollinearity 37

5 Discussion 37
5.1 Evaluation of final model 37
5.2 Impact of the regressor variables 38
5.2.1 Rating 38
5.2.2 Genre 39
5.2.3 Sequel 39
5.2.4 RottenTomatoesScore 39
5.2.5 IMDbScore 40
5.2.6 DirectorScore 40
5.2.7 CreativeType 40
5.2.8 Source 40
5.2.9 ProductionMethod 41
5.2.10 Confidence intervals of the regressors 41
5.3 Limitations of model and approach 42
5.3.1 Budget estimations 42
5.3.2 Marketing 42
5.3.3 Casting 42
5.3.4 Initial model formulation 43
5.3.5 Limitation due to scope 43
5.4 Conclusion 43
6 Porter’s five forces analysis

6.1 Introduction ................................................. 44
6.2 Scope ......................................................... 44
6.3 Problem statement ............................................. 45
6.4 Method ........................................................ 45
  6.4.1 Types of movies ........................................ 45
  6.4.2 Choice of Porter’s five forces analysis ................. 46
6.5 Theory ........................................................ 46
6.6 Five forces applied on the cinematic theater industry ... 47
  6.6.1 Threat of new entrants ................................. 47
  6.6.2 Threat of substitutes .................................... 48
  6.6.3 Bargaining power of customers ......................... 49
  6.6.4 Bargaining power of suppliers ......................... 49
  6.6.5 Industry rivalry .......................................... 50
6.7 Discussion ..................................................... 50
6.8 Conclusion .................................................... 52
6.9 Limitation of approach ....................................... 52
1 Introduction

1.1 Background

Motion pictures have been a staple in the modern society ever since the emergence of dedicated movie theaters at the beginning of the 20th century. The global box office revenue is forecast to increase from about $38 billion in 2016 to nearly $50 billion in 2020[1].

Creating a movie is an intricate process with multiple stages and hurdles. Production requires a lot of capital upfront and a long wait until any returns materialize. Movie production is consequently a risky venture since there exists a lot of uncertainty on whether a movie will be profitable or not. This is especially a problem for big-budget movies since there is a lot of capital at stake. The production of a movie can be broken down into five important stages[2]:

- **Development**
  The stage where the movie idea is determined, necessary rights are acquired, the screenplay is written and where the financing for the movie is obtained.

- **Pre-production**
  Preparations before shooting are made such as hiring cast and crew, locations to shoot in are located and constructing necessary sets and props.

- **Production**
  The actual shooting of the movie where raw footage and the additional sound is recorded.

- **Post-production**
  The recorded footage and sound is edited and combined with visual effects and music into the final product.

- **Distribution**
  The movie is then marketed, distributed and released to cinemas or other media.

A select number of large studios are attempting to capture the audience’s attention. 89% of total box office revenue in 2017 was accounted for by 7 different studios[3]. But not all studios are successful, Sony Pictures division for instance reported a $913 million loss in the quarter ending in Dec 31 2016[4]. Viacom reported a $364 million loss in 2016[5].

The risk associated with creating blockbusters could be an explanation to the rise of movie sequels, remakes and adaptations that can be seen in today’s movie industry where these kinds of...
movies are a safer investment since they already have an established audience. Multiple different investment decisions apart from making sequels exists in the different production phases. To which screenplay should the rights be bought? Should A-list actors be cast? When should the movie be released? How should the movie be distributed? Examining how these types of decisions affect profitability is the goal of this thesis.

1.2 Purpose and motivation

An aim of this project is to use regression analysis to gain insight regarding what influences the profitability of a movie. If this thesis is able to produce a satisfactory model for profitability then it would be able to mitigate some of this risk associated with investment decisions in movie production. A movie is still something that is artistic in its nature so believing that a mathematical model can rule is nonsensical but knowing how a certain decision during the production of a movie will shift the profitability horizon could still be of value.

How a movie should be distributed with regard to profitability is also a main factor which the thesis will examine. This will specifically be done with a porters five forces analysis focused on the movie theatre industry. The reasoning behind this analysis is to complement the regression analysis, which target intrinsic values of a movie and its effect on profitability, by targeting distributions effect on profitability.

1.3 Disposition

Since this thesis consists of two parts (the regression analysis and Porter’s five forces analysis) they will have their own scope and problem statement. The scope and problem statement for the regression analysis will be presented below and the scope and problem statement for Porter’s five forces analysis will be presented later in section 6.

1.4 Scope

The thesis will focus on motion pictures with at least a cinematic release and an estimated budget greater than $5 million. As a result, the thesis will focus on movies produced by relatively established movie studios which is our goal. The thesis will investigate movie profitability with respect to multiple variables and movies released within 2015-2017 will be used for this part. This restriction is made to ensure that the model is not affected by changes, such as trend shifts, dependent on greater time differences. It would not make sense to compare movies from this year to movies released in the eighties since the influencing factors have probably changed since then. What kind of potential profitability influencing factors to examine and the scope associated with those decisions will be discussed in section 3.1.

---

1.5 Problem Statement

- Is it possible to build a model with multiple linear regression to predict and/or understand movie profitability?
- Which factors influence profitability the most?

2 Mathematical Theory

2.1 Multiple linear regression

Regression analysis is a statistical technique for explaining and modeling the relationship between different variables. The general idea of linear regression is that you can model the relationship between a response variable $y$ and another variable $x$ called the regressor variable. The response can thereby be explained by the regressor variable since it is dependent on it. The multiple linear regression model is just a linear regression model with multiple regressor variables $x_i$ that has a linear relationship with the response variable $y$. The relationship can be described by the following equation:

$$ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon $$  (3.1)

Where $y$ is the response variable which is dependent on $k$ regressor variables $x$ and the error $\epsilon$. The parameters $\beta_j$, $j = 0, 1, 2, 3, \ldots, k$ are called the regression coefficients where $\beta_0$ is the intercept. The coefficients can be determined by using collected data and the method of least squares. A more convenient way of displaying the model is to use matrix notations which transform equation (3.1) to the following equation:

$$ y = X\beta + \epsilon $$  (3.2)

where

$$ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} $$

$$ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} $$
2.2 Ordinary Least Squares

The parameter $\beta$ in the equation of the multiple linear regression model is unknown and must be estimated. Ordinary least squares (OLS) is a method that can be used to estimate the regression coefficient $\beta$. OLS estimates the coefficients using sample data where $n$ observations will give different data points of the following kind which is called the sample regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \quad (3.3)$$

$i = 1, 2, ..., n$

OLS estimates the coefficients so that the sum of the squares of the differences between the observations of the dependent variable $y_i$ and the predicted values from the linear model in equation (3.1) is at a minimum.

When using the matrix notations introduced in equation (3.2) the goal is to find the vector of least-squares estimators, $\hat{\beta}$, that minimizes the least-square function:

$$S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon' \epsilon = (y - X\beta)'(y - X\beta) \quad (3.4)$$

Which gives the least squares normal equations.

$$X'X\hat{\beta} = X'y \quad (3.5)$$

Which when solved gives the least squares estimator of $\beta$.

$$\hat{\beta} = (X'X)^{-1}X'y \quad (3.6)$$

provided that the inverse matrix $(X'X)^{-1}$ exists.

2.3 Assumptions

The assumptions made when creating multiple linear regression models are the following:

1. The relationship between the response variable and the regressor variables is linear, at least approximately.
2. The error term $\epsilon$ has zero mean.
3. The error term $\epsilon$ has constant variance $\sigma^2$.
4. The errors are uncorrelated.
5. The errors are normally distributed.

---

9 ibid p. 71
10 ibid p. 72
11 ibid p. 129
The linear regression model assumptions must be examined and must hold before using the regression model. A number of graphical methods exist to identify any violations of the assumptions and there also exists methods to handle these violations such as different transforms of either the response variable or regressor variables. The methods used in this thesis will be presented below.

2.4 Dealing with violation of assumptions

2.4.1 Residuals vs fitted

A plot of residuals versus the corresponding fitted values can be useful to identify violations of the linear regression assumptions. Violations can be identified by searching for patterns in the plot. The desired pattern is a horizontal band which means that the variance is constant and is called homoscedasticity. An unwanted pattern could be funnel shapes which imply that the variance is not constant, which is called heteroscedasticity, or non-linear shapes which imply that the relationship between response variables and regressor variables are not linear. Transforms of response or regressor variables can be used to ensure that the assumptions hold.

2.4.2 Residuals vs regressors plot

A plot of residuals versus regressor variables is also helpful to identify violations of the assumptions. To identify violations patterns are searched for as with the plot of residuals versus the fitted values. Transforms of the regressor variables can be the solution to this problem.

2.4.3 Q-Q plot

The Q-Q plot, which stands for the quantile-quantile plot, is a graphical method which can be used to determine if some data comes from some theoretical distribution. This is achieved by plotting two sets of quantiles against each other and if the quantiles truly come from the same distribution the plotted points should form a line. This is useful for checking if the assumption of normally distributed residuals with mean zero holds. The normal Q-Q plot is created by plotting the standardized residuals against the theoretical quantiles which should form a line if truly normally distributed.

2.4.4 Histogram of residuals

Another way to check if the normality assumptions with mean zero holds is to investigate a histogram of the residuals. The histogram should follow a normal distribution with mean zero.

2.4.5 Scale-location plot

The scale-location plot can be useful for identifying violations against the assumption of constant variance. It is a plot of the square root of the standardized residuals versus the fitted values and shows if residuals are equally spread along the range of the regressors. The desired result is a horizontal line with randomly equally spread points.

\[\text{ibid p. 139}\]
\[\text{ibid p. 141}\]
2.4.6 Box-Cox Transform

The Box-Cox transform can be used to correct non-normality or non-constant variance\(^\text{14}\) It uses the power transformation \(y^\lambda\) where \(\lambda\) is a parameter which needs to be estimated. A problem arises when \(\lambda = 0\) since then the transformation will be worthless. A fix for this is to use:

\[
\frac{(y^\lambda - 1)}{\lambda} \quad \text{when} \quad \lambda \neq 0
\]

and \(\log(y)\) when \(\lambda = 0\)

The \(\lambda\) can be estimated computationally by minimizing the residual sum of squares from the fitted model \(SS_{Res}(\lambda)\). The wanted \(\lambda\) is the one which minimizes this function.

2.5 Multicollinearity

The problem of Multicollinearity arises when there exist near-linear dependencies among the regressor variables and will have a serious impact on the least-square estimators\(^\text{15}\). A strong multicollinearity between regressor will lead to large variances and covariances for the least-square estimators of the regression coefficients\(^\text{16}\).

2.5.1 VIF

The variance inflation factor can be used when identifying multicollinearity. It is based on the matrix \(C = (X'X)^{-1}\) where \(j\)th diagonal element of \(C\) can be written as \(C_{jj} = (1 - R_j^2)^{-1}\), where \(R_j^2\) is the coefficient of determination\(^\text{17}\). If \(x_j\) is linearly dependent on some of the remaining regressors, \(C_{jj}\) will be large and since the variance of the \(j\)th regression coefficient is \(C_{jj}\sigma^2\), \(C_{jj}\) can be viewed as a factor the least-square estimators variance is increased with because of the near-linear dependencies with other regressors. So \(VIF_j\) is therefore the following.

\[
VIF_j = C_{jj} = (1 - R_j^2)^{-1}
\]

Where a VIF which exceeds 5 or 10 is considered to be an indication of that multicollinearity affects the regression coefficients negatively\(^\text{18}\).

2.6 Detecting leverage and influential observations

When creating a multiple linear regression model it is important to investigate if leverage or influential observations exist within the chosen data-set. The difference between leverage points and influential observations is that leverage points lie approximately along the regression line but with abnormal x-values while influential observations depart from the regression line.

Influential points have a noticeable impact on the regression model and the model will be drawn

\(^{14}\)ibid p. 182
\(^{15}\)ibid p. 285
\(^{16}\)ibid p. 289
\(^{17}\)ibid p. 296
\(^{18}\)ibid p. 296
towards these points. This is problematic since a model could be severely affected by a relatively few amount of points. Influential points that have a bad impact on the model should, therefore, be considered for removal.

While not all points with unusual x-values are influential they can potentially play an important role when determining the properties of the regression model. Remote leverage points could have a disproportionate impact on the model parameters and should, therefore, be identified and handled.

Diagnostics to detect influential points is presented below.

### 2.6.1 Cook’s distance

The Cook’s distance is a diagnostic tool for detecting influential and leverage points and does this by measuring the squared distance between the least-squared estimate $\hat{\beta}$ based on all n points in the dataset and the estimate $\hat{\beta}^{(i)}$ which is obtained by deleting the ith point. The Cook’s distance can be expressed as

$$D_i(X'X, pMS_{res}) = \frac{(\hat{\beta}^{(i)} - \hat{\beta})'X'X(\hat{\beta}^{(i)} - \hat{\beta})}{pMS_{res}}, \quad i = 1, 2, ..., n \quad (3.7)$$

where $MS_{res}$ is the mean square of the residuals and $p$ is the number of regressors. Observations with large $D_i$ values will have a noticeable influence on the least-square estimators and needs to be handled some way. This thesis will classify $D_i$ values as large when they are greater than 1.

### 2.6.2 Residuals vs Leverage plot

The residuals versus leverage plot is a useful plot to detect influential observations. Data points with a high influence will influence the model greatly and the model would not be the same if this data point is eliminated. An observation which is not influential would not change the model much if eliminated. The residuals vs leverage plot will contain dashed lines representing certain values of Cook’s distance. Influential observations can be identified since these will be outside of the dashed lines representing a Cook’s distance equal to 1.

### 2.7 Hypothesis testing

#### 2.7.1 F-Statistic

The F-test is used to determine if two populations variances are equal or not and does this by comparing the variances between the populations and within the populations. This is useful when determining whether the linear regression model in question provides a better fit to the data than the model with no regressor variables. The two hypothesis of the F-test for multiple regression model are the following.

\[ \text{ibid p. 211} \]
\[ \text{ibid p. 212} \]
\[ \text{ibid p. 215} \]
\[ \text{ibid p. 84} \]
• The null hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
  Which means that the model with no regressor variables fits the data as good as the model in question.

• The alternative hypothesis: $H_1 : \beta_j \neq 0$ for at least one $j$
  Which means that the model in question fits the data better than the model with no regressor variables.

A rejection of the null hypothesis implies that at least one of the regressors are statistically significant. The F-statistic which can be acquired through an ANOVA-table can be used when deciding to reject or accept the null hypothesis and is defined as the following.

$$F_0 = \frac{\text{variance between populations}}{\text{variance within populations}} = \frac{MS_R}{MS_{res}}$$

Where $MS_R$ is the mean square due to regression and $MS_{res}$ is the mean square due to the residuals.

If the null hypothesis holds this division would be equal to one since the variances between and within the populations would be the same and the model in question would be as good as the model with no variables. A large positive F-statistic that is far away from one would, therefore, be the desired result.

The P-value which also is provided through an ANOVA-table is the probability of getting the acquired F-statistic while the null hypothesis is true. Because of this, the corresponding P-value must always be used alongside the F-statistic. If the P-value is lower than the set significance level the null hypothesis can be rejected.

2.7.2 t-test

A t-test is used to test the significance of an individual regressor coefficient and is based on the t distribution. Important to note is that this test the significance of a coefficient while the remaining regressors are included in the model. The hypotheses are as follows

$H_0 : \beta_j = 0, H_1 : \beta_j \neq 0$

Test statistic $T_0$ is computed as $T_0 = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

The null hypothesis fails to be rejected if

$-t_{n-2} < T_0 < t_{n-2}$

2.8 Variable selection and model selection

2.8.1 All possible regression

All possible regression is a method to reduce a model to possibly find a better model. The method generates all possible regression models based on the available regressors and then compares these models. There are several different criteria that can be used to compare the models which means that it is seldom possible to arrive at a single final model.
All possible regression is preferable compared to types of stepwise regression since it is more thorough. A downside to the method is the computational load since there are $2^k$ different models to examine.

### 2.8.2 Cross-validation

Cross-validation is a method to examine the predictive power of a model. The principle behind cross-validation is that the model is fitted to a specific part of the dataset, this part is called the training set and then tested on the rest of the data, called the test set. k-fold cross validation means that the data is randomly split into k folds. One fold is used as a test set and the remaining k-1 folds are used as training set. This process is then repeated for each fold. The k different results can then be averaged to generate a final result. This thesis will use $k=10$.

### 2.8.3 $R^2$ and adjusted $R^2$

$R^2$, the coefficient of multiple determination, measures the proportion of variance in the response variable that is explained by the regressors and is a suitable criteria for model selection. Commonly used as a measure of goodness of fit. $R^2$ is in this thesis defined as

$$R^2 = 1 - \frac{SS_{Res}(p)}{SS_T}$$

A problem with $R^2$ is that it is always possible to make it grow by adding new regressors. This can make it harder to interpret the $R^2$ value and lead to overfitting. Adjusted $R^2(\bar{R}^2)$ aims to solve this problem by penalizing the inclusion of new regressors. $\bar{R}^2$ is defined as

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

### 2.8.4 Mallows’s $C_p$ statistic

$C_p$ is another criterion used to measure goodness of fit and is suitable as a criteria for model selection. It is defined as

$$C_p = \frac{SS_{Res}(p)}{\sigma^2} - n + 2p$$

It can be shown that the expected value of $C_p$ given that there is no bias in the model of size $p$ is equal to $p$. Models with small bias will consequently be near $p$ and models with larger bias will be larger than $p$. Small values of $p$ are generally better.

### 2.8.5 AIC

Akaike information criterion (AIC) is a measure that can be used for model selection, it provides only information regarding relative quality between models. AIC both rewards goodness of fit and penalizes the inclusion of new regressors, thus preventing overfitting. When comparing models the model with the lowest AIC value will be the model that minimizes estimated information loss. It is calculated as

---

23ibid p. 335
\[ AIC = -2\ln(L) + 2p \]
where \( L \) is the maximum likelihood function for the model and \( p \) is the number of regressors.

2.9 Dummy variables

The regressor variables used in regression analysis can be of two different types, quantitative or qualitative. Quantitative variables have a scale of measurement where a budget of 1 000 000 is larger than a budget of 1000 for example. Qualitative variables which are also called categorical variables lack this sense of scale and it is not possible to determine any kind of meaningful order between the different levels within the categorical variable. It is not possible to order genres in the same way that budgets could be ordered.

A method to include qualitative variables into the regression model is to use dummy variables. For a qualitative variable, say genre, each level of it, say horror or action, will become a dummy variable expect for one level which will be a reference. Each dummy variable can be assigned a 1 or 0 depending on if the movie in question is classified as the genre that the dummy variable represents. A horror movie would, therefore, score a 1 on the dummy variable for horror movies but a 0 for the one for action.

3 Method

3.1 Data collection

A dataset has been given to us from Opus Data which have data concerning the types of variables that are supposed to be investigated in this thesis. Data has also been manually sourced from RottenTomatoes and IMDb.

3.2 Software

R was used in this thesis to perform the data processing and regression analysis.

3.3 Dataset from Opus Data

The dataset from Opus Data contained 322 observations within the time span of 2015-2017. Each observation had 13 variables associated with it, the variables are presented below.

Variables in Opus Data dataset

- **movie name**
  Name of the movie

- **production year**
  Numerical variable.

\[ \text{ibid p. 260} \]
• **movie odid**
  Numerical identifier for the movie.

• **production budget**
  Numerical variable. Budget is showcased in dollars.

• **domestic box office**
  Numerical variable. The Domestic Market is defined as the North American movie territory (consisting of the United States, Canada, Puerto Rico and Guam).

• **international box office**
  Numerical variable. Revenue from ticket sales outside of the domestic market.

• **rating**
  Categorical variable. Motion Picture of America(MPAA) film rating. Used mainly in the USA.

• **creative type**
  Categorical variable highlighting if the movie is for instance a science fiction movie or a fantasy movie.

• **source**
  Categorical variable showing what the movie is based on, for instance if it has an original screenplay or if it’s based on a book.

• **production method**
  Categorical variable showing if the movie for instance is a live action movie or an animated movie.

• **genre**
  Categorical variable showing which genre the movie belongs in.

• **sequel**
  Binary variable. 1 if the movie is a sequel, 0 otherwise.

• **running time**
  Numerical variable. Shows how long the movie is in minutes.

### 3.4 Preprocessing of Opus Dataset

One major error in the dataset was found. A movie in the dataset had a running time of 0 minutes. This was consequently deemed as an input error and the actual running time, 85 minutes, was sourced from IMDb and manually inserted in the dataset.

There existed some variable categories that only appeared in a single observation, these types of unique categories will cause problems if a regression analysis is performed. The problem was found in the rating variable where only one observation had a rating *G*. Instead of removing this observation this movie was treated as if it had a *PG* rating. Source also contained observations with unique categories. These movies were given a source of NaN, i.e. these movies did not belong to any of the major source categories.
3.5 Creation of new variables

The dataset from Opus Data contained no information regarding the director of the movie nor public and critical reception of the movies. Three new variables were created to mitigate this shortcoming.

The first new variable is Rotten_Tomatoes_Score. This numerical variable is the movie's Tomatometer score from Rotten Tomatoes. This Tomatometer score represents the percentage of professional critic reviews that are positive towards the movie. Some movies did not have a Tomatometer score, these movies were given a score of 50% meaning that half of every potential reviewer was assumed to be positive towards the movie.

The second new variable is IMDb_Score. This numerical variable is the movie's IMDb rating score. IMDb ratings score is a weighted average of IMDb users ratings. Any user on the IMDb website is eligible to rate a movie.

The third new variable is DirectorScore. This variable is a categorical variable containing 4 different categories A, B, C and D. A director score of A means that the director for the given movie has during their lifetime generated more than 1.5 billion $ in aggregated box office revenue. B represents aggregated box office revenue in the range of 1.5-1 billion $, C the range of 1-0.5 billion $ and D <0.5 billion $. The aggregated box office revenue is calculated as the sum of all the box office revenues for each movie that the director has directed. The director for each movie was manually sourced from imdb.com and the aggregated box office revenue was sourced from boxofficemojo.com. The goal of the variable was to highlight the degree of experience for a movie's director. Aggregated box office was chosen as an indirect measure of experience since a large aggregated box office would indicate a long directorial career and/or financially successful movies.

3.6 Initial transformation of variables

The numerical variable running time was transformed into a categorical variable with three categories a, b and c. a represents a movie with running time over 130 minutes, b represents a movie with running time between 110 and 130 minutes and c represents a movie with a running time below 110 minutes.

This transformation decision was based on a discussion by Randy Olson which showed that the average feature film length for the 25 most popular movies for each year between 2000 and 2013 had a 95% confidence interval that approximately contained the movies with running times from 110 to 130 minutes. This means that the categorical transformation highlights whether or not a movie's running time is normal or if it is short or long.

3.7 Choice of response variable

This thesis is mainly interested in how profitability is affected by certain factors. Due to this, the response variable is chosen to be the return percentage that each movie achieves.

\[ y = \frac{\text{domestic box office} + \text{international box office} - \text{production budget}}{\text{production budget}} \]

3.8 Choice of regressors

The regression model will include rating, creative type, source, production method, genre, sequel, running time (as categorical variables), IMDb score, rotten tomatoes score and director score as regressors. A Categorical variable with k categories will be coded with k-1 dummy variables representing the k-1 categories, the final category is held as a reference category. Binary variables are coded as a dummy variable. Numerical variables will be coded as continuous variables.

3.9 Initial model

The initial model with all variables included is presented below.

\[ \text{Profitability} = \text{Rating} + \text{Genre} + \text{Sequel} + \text{RunningTime} + \text{Rotten Tomatoes Score} + \text{IMDb Score} + \text{Director Score} + \text{Budget} + \text{Creative Type} + \text{Source} + \text{Production Method} \]

Where the categorical variables have been divided into dummy variables as presented below with corresponding reference variable.

- **Rating**
  - PG, PG_13, and R with reference Not Rated

- **Genre**
  - Action, Adventure, Comedy, Drama, Horror, Musical, Romantic Comedy, Thriller/Suspense and Western with reference Black Comedy

- **Sequel**
  - The categorical variable Sequel only consists of the dummy variable sequel where the references are the movies that are not sequels.

- **RunningTime**
  - b and c with reference a, where the meaning of each variable has been presented in section 4.6

- **DirectorScore**
  - A, B and C with reference D, where the meaning of each variable has been presented in section 4.5

- **CreativeType**
  - Historical Fiction, Dramatization, Fantasy, Kids Fiction, Science Fiction and Super Hero with reference Contemporary Fiction
• **Source**
  Comic, Factual Book, Fictional Book, Game, RealLife Events, Religious Texts, Toy, TV, NaN, Original Screenplay, Remake and Spin Off (where NaN means that no source could be found for the movie) with reference FolkTale

• **ProductionMethod**
  Digital Animation, Live Action and Animation/Live Action with reference StopMotion Animation

The result presented below is for the initial model which can be seen above.
| Coefficients       | Estimate  | Std.Error  | t value | Pr(>|t|)  |
|--------------------|-----------|------------|---------|----------|
| (Intercept)        | 3.898119  | 5.800731   | 0.672   | 0.502137 |
| PG                 | 1.836643  | 1.001410   | 1.834   | 0.067711 .|
| PG-13              | 1.156377  | 0.858991   | 1.346   | 0.179329 |
| R                  | 0.632033  | 0.842452   | 0.750   | 0.453748 |
| Action             | 1.235737  | 1.309392   | 0.971   | 0.332534 |
| Adventure          | 1.552084  | 1.370883   | 1.132   | 0.258532 |
| Comedy             | 1.078496  | 1.287291   | 0.838   | 0.402859 |
| Drama              | 0.759524  | 1.286222   | 0.589   | 0.556371 |
| Horror             | 5.423540  | 1.383652   | 3.920   | 0.000112 ***|
| Musical            | 8.189243  | 2.079921   | 3.937   | 0.000104 ***|
| Romantic_Comedy    | 1.749185  | 1.628423   | 1.074   | 0.283680 |
| Thriller/Suspense  | 0.698192  | 1.291606   | 0.541   | 0.589241 |
| Western            | -0.340217 | 1.832720   | -0.186  | 0.852787 |
| Sequel             | 1.534141  | 0.418383   | 3.663   | 0.000298 ***|
| b                  | -0.005416 | 0.386724   | -0.141  | 0.888837 |
| c                  | -0.023020 | 0.594198   | -0.039  | 0.969124 |
| Budget             | -0.815640 | 0.294290   | -2.666  | 0.008368 **|
| Rotten_Tomatoes_Score | 0.024952  | 0.008863   | 2.815   | 0.005219 **|
| IMDb_Score         | 0.061656  | 0.027786   | 2.219   | 0.027293 * |
| A                  | 0.106156  | 0.778103   | 0.136   | 0.891580 |
| B                  | 0.424383  | 0.793886   | 0.535   | 0.593377 |
| C                  | 0.252379  | 0.521024   | 0.484   | 0.628489 |
| Historical_Fiction | -0.722422 | 0.592196   | -1.220  | 0.223646 |
| Dramatization      | -0.847871 | 1.057981   | -0.801  | 0.423580 |
| Fantasy            | -0.402737 | 0.650768   | -0.626  | 0.531935 |
| Kids_Fiction       | -0.783030 | 1.387664   | -0.564  | 0.573073 |
| Science_Fiction    | -0.088514 | 0.566051   | -0.156  | 0.875854 |
| Super_Hero         | 1.649517  | 1.337989   | 1.237   | 0.217236 |
| Comic              | -0.038081 | 1.746843   | -0.218  | 0.827585 |
| Factual_Book       | 0.685913  | 1.824541   | 0.376   | 0.707249 |
| Fictional_Book     | 0.213300  | 1.458384   | 0.146   | 0.883824 |
| Game               | 1.576121  | 1.859891   | 0.847   | 0.397484 |
| RealLife_Events    | -0.806244 | 1.801242   | -0.481  | 0.630954 |
| Religious_Texts    | -0.073325 | 2.399984   | -0.031  | 0.975648 |
| Toy                | -1.238599 | 2.132597   | -0.581  | 0.561848 |
| TV                 | -0.166935 | 1.703328   | -0.098  | 0.921998 |
| NaN                | 0.009436  | 2.104354   | 0.004   | 0.996425 |
| Original_Screenplay| 0.595533  | 1.442913   | 0.413   | 0.680122 |
| Remake             | 1.131599  | 1.679801   | 0.674   | 0.501091 |
| Spin_Off           | 2.489375  | 2.262762   | 1.100   | 0.272216 |
| Digital_Animation  | 4.368805  | 2.287014   | 1.910   | 0.057124 .|
| Live_Action        | 3.960267  | 2.282161   | 1.735   | 0.083789 .|
| Animation/Live_Action | 2.572488 | 2.286218   | 1.125   | 0.0261465 |

Page 23
Multiple R-squared     0.3915
Adjusted R-squared     0.2999
p-value                9.926e-14

Table 1: ANOVA-table for the initial model

<table>
<thead>
<tr>
<th>Regressors</th>
<th>Mean sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>17.226</td>
<td>2.4144</td>
<td>0.066862</td>
</tr>
<tr>
<td>Genre</td>
<td>65.101</td>
<td>9.1246</td>
<td>4.168e-12 ***</td>
</tr>
<tr>
<td>Sequel</td>
<td>76.612</td>
<td>10.7379</td>
<td>0.001182 **</td>
</tr>
<tr>
<td>Running_Var</td>
<td>5.592</td>
<td>0.7838</td>
<td>0.457679</td>
</tr>
<tr>
<td>Rotten_Tomatoes_Score</td>
<td>267.705</td>
<td>37.5216</td>
<td>3.066e-09 ***</td>
</tr>
<tr>
<td>IMDb_Score</td>
<td>24.729</td>
<td>3.4660</td>
<td>0.063696 .</td>
</tr>
<tr>
<td>Director_Score</td>
<td>2.812</td>
<td>0.3941</td>
<td>0.757336</td>
</tr>
<tr>
<td>Budget</td>
<td>45.938</td>
<td>6.4387</td>
<td>0.011711 *</td>
</tr>
<tr>
<td>Creative_Type</td>
<td>12.855</td>
<td>1.8018</td>
<td>0.098711 .</td>
</tr>
<tr>
<td>Source</td>
<td>5.682</td>
<td>0.7963</td>
<td>0.654229</td>
</tr>
<tr>
<td>Production_Method</td>
<td>13.708</td>
<td>1.9213</td>
<td>0.126308</td>
</tr>
</tbody>
</table>

Where the significance codes are: 0 (***), 0.001 (**), 0.01 (*), 0.05 (.)

3.9.1 Residuals plotted against the fitted values

![Figure 1: The residuals plotted against the fitted line for the initial model](image)
There exists no problem with the line but the shape of the plotted residuals is problematic since the variance is not constant. The shape is of an outward-opening-funnel and implies that the variance is an increasing function of the response variable. Heteroscedasticity exists which means that the third assumption is violated. To deal with this problem some kind of transformation of either the response variable or the regressor variable should be performed.

### 3.9.2 The residuals plotted against the regressor variables

To ensure that the assumption of constant variance and mean zero for the error $\epsilon$ holds the plots of the residuals against the regressors has been investigated. The only noticeable problem that arose was with the regressor variable budget.

![Figure 2: Residuals plotted against budget for the initial model](image)

The residuals plotted against the regressor variable budget displays a problematic pattern, even though the assumption that the mean of the residuals is zero holds since the variance of the residuals can’t be considered to be constant. The pattern of an inward-opening-funnel is an indication of that the variance of the residual increases as the regressor variable decreases. A transformation of this regressor variable is necessary to make sure that the assumption of constant variance holds.

---

3.9.3 Normal Q-Q plot

The residuals of the initial models seem to be somewhat normally distributed with mean zero. There exists a noticeable departure from the line in the right edge which could be problematic. Small departures from the line don’t impact the model heavily but major departures can cause serious problems since the t or F statistic and confidence or predictions intervals depend on the assumption that the residuals are normally distributed.

3.9.4 Histogram of residuals

Figure 3: A Q-Q plot for the initial model

Figure 4: A histogram of the residual for the initial model

\[^{27}\text{ibid p. 136}\]
The residuals seem to be somewhat normally distributed, as they were when examining the Q-Q plot, with the mean zero. The histogram seems somewhat skewed towards one side and is not entirely equally distributed around zero.

3.9.5 Scale-Location plot

![Scale-Location plot](image)

Figure 5: The Scale-Location plot for the initial model

The Scale-Location plot of the initial model is not of the desired type. The desired plot is a horizontal line with randomly equally spread points which is not the case here.
3.9.6 Residual vs leverage plot

![Residual vs Leverage Plot](image)

Figure 6: A plot of the residual vs leverage for the initial model

While examining the plot which can be used for identifying leverage and influential points no such points were detected since all the observations lies within the two dashed lines.

3.9.7 Multicollinearity

<table>
<thead>
<tr>
<th>Regressor</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>5.078326</td>
</tr>
<tr>
<td>Genre</td>
<td>15.726641</td>
</tr>
<tr>
<td>Sequel</td>
<td>1.393611</td>
</tr>
<tr>
<td>Running_Var</td>
<td>2.116239</td>
</tr>
<tr>
<td>Rotten_Tomatoes_Score</td>
<td>2.808200</td>
</tr>
<tr>
<td>IMDb_Score</td>
<td>3.157547</td>
</tr>
<tr>
<td>Director_Score</td>
<td>2.052309</td>
</tr>
<tr>
<td>Budget</td>
<td>3.484704</td>
</tr>
<tr>
<td>Creative_Type</td>
<td>338.960312</td>
</tr>
<tr>
<td>Source</td>
<td>70.814768</td>
</tr>
<tr>
<td>Production_Method</td>
<td>15.469296</td>
</tr>
</tbody>
</table>
The problem of multicollinearity exists in the initial model since VIF values which exceeds 5 or 10 is present. Some variables need to be eliminated to solve this problem.

### 3.10 Transformation of the model

A box-cox transformation was used to solve the issues of non-normality as seen in the Q-Q plot and the heteroskedasticity as seen in the various residual plots. The lambda that minimizes $SS_{Res}(\lambda)$ is presented below.

$$\lambda = 0.34343434$$

Transforming $y$ with $\lambda$ led to a satisfactory Q-Q plot which did not show any major signs of non-normality. The funnel pattern seen in the residual plots was also reduced greatly meaning that the assumption of constant variance approximately holds after the transformation.

To amend the problematic pattern in the regressor variable budget vs residuals plot the regressor variable budget was log-transformed. This had a positive effect on the pattern.

These transformations eliminated the aforementioned issues of violated assumptions. The effects of these transformations can be seen in section 4 where the relevant plots are presented for the final model.

### 3.11 Reduction of the model

Two approaches to model reduction were taken. The first approach was based on maintaining the integrity of categorical variables, i.e. a categorical variable was either included or not included. This approach reduced the number of regressors to examine and made it is possible to use all possible regression along with criteria such as adjusted $R^2$, Mallows’s Cp and AIC.

The second approach was to allow manipulation of categorical variables, i.e. every single dummy variable could be included or not included. This approach was considerably more computationally complex compared to the first but meant more model flexibility. To combat this problem of computational load all possible regression was performed with $SS_{Res}$ as criteria, the models with lowest $SS_{Res}$ for each model size was then further analyzed. From the generated models, one for each model size, the best was picked with regard to adjusted $R^2$.

K-fold cross validation was finally used to create a summarizing comparative criterion for the models created by the first and second approach.

#### 3.11.1 Models generated by the first approach

Six models were picked based on adjusted $R^2$, these models are denoted models 1 to 6. Several models were picked since these performed similarly. Six models were also picked based on Mallows’s Cp, these models were denoted models 7 to 12. Three models were picked based on AIC, these models were denoted 13 to 15.
3.11.2 Models generated by the second approach

Among the models generated by the SS_{Res} ranking the six best based on adjusted $R^2$ were picked, these models were denoted 16 to 21.

3.11.3 Cross validation results

K-fold cross-validation was performed for $k=10$ and models were compared with regards to MSE. All the models from the second approach outperform the models from the first approach in the cross-validation.

Table 3: Cross validation results

<table>
<thead>
<tr>
<th></th>
<th>17</th>
<th>19</th>
<th>16</th>
<th>18</th>
<th>21</th>
<th>20</th>
<th>15</th>
<th>1</th>
<th>13</th>
<th>14</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>1.3845</td>
<td>1.3902</td>
<td>1.3903</td>
<td>1.3922</td>
<td>1.4055</td>
<td>1.4095</td>
<td>1.4622</td>
<td>1.4668</td>
<td>1.4668</td>
<td>1.4738</td>
<td>1.4915</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>1.4956</td>
<td>1.5074</td>
<td>1.5108</td>
<td>1.5166</td>
<td>1.5271</td>
<td>1.6156</td>
<td>1.6234</td>
<td>1.6422</td>
<td>1.6446</td>
<td>1.6555</td>
<td></td>
</tr>
</tbody>
</table>

3.11.4 Choice, verification and transformation of final model

Since the top six models based on cross-validation performs similarly the approach in this thesis was to pick the model that is as small as possible and has as low VIF values as possible. This lead to the choice of model 21.

The residual plots for model 21 does not indicate any major violations of the initial assumptions. On the other hand, model 21 has problems with multicollinearity, there are as seen three VIF values over 5 indicating multicollinearity issues.

To combat the multicollinearity the dummy variable R is removed, i.e. it becomes part of the reference. This resulted in a model with no VIF values over 5. The adjusted $R^2$ shrank from 0.3723 to 0.3691. In effect of the marginal decrease in adjusted $R^2$ and the fact that it is unknown whether or not the present multicollinearity will hold in the future, model 21 with the dummy variable R removed is chosen as the final model.

4 Result

4.1 Final model

The reduced and final model is presented below.

$$\text{Profitability} = \text{Rating} + \text{Genre} + \text{Sequel} + \text{RottenTomatoesScore} + \text{IMDbScore} + \text{DirectorScore} + \text{CreativeType} + \text{Source} + \text{ProductionMethod}$$

Where the categorical variables have been divided into dummy variables as presented below. Since several dummy variables have been removed the references consist of several different categories. For each category, the reference consists of those categories not specified in the model.
• **Rating**  
PG and PG_13

• **Genre**  
Adventure, Drama, Horror and Musical

• **Sequel**  
The categorical variable Sequel only consists of the dummy variable sequel where the references are the movies that are not sequels.

• **DirectorScore**  
C where the meaning of each variable has been presented in section 3.5

• **CreativeType**  
Historical_Fiction, Dramatization and Science_Fiction

• **Source**  
Factual_Book, Game, Religious_Texts, Original_Screenplay and Spin_Off

• **ProductionMethod**  
Live_Action

The result presented below is for the final model which can be seen above.

<p>| Coefficients | Estimate  | Std.Error | t value | Pr(&gt;|t|) |
|--------------|-----------|-----------|---------|---------|
| (Intercept)  | -1.700357 | 0.642108  | -2.648  | 0.008520 ** |
| PG           | 1.002975  | 0.275367  | 3.642   | 0.000318 *** |
| PG-13        | 0.369769  | 0.143060  | 2.585   | 0.010215 *  |
| Adventure    | -0.349524 | 0.248091  | -1.409  | 0.159907   |
| Drama        | -0.317009 | 0.194189  | -1.632  | 0.103621   |
| Horror       | 1.411846  | 0.296416  | 4.763   | 2.97e-06 *** |
| Musical      | 1.608555  | 0.677956  | 2.373   | 0.018287 *  |
| C            | 0.250866  | 0.200419  | 1.252   | 0.211645   |
| Historical_Fiction | -0.900884 | 0.223523 | -4.030  | 7.05e-05 *** |
| Dramatization | -1.016849 | 0.249081 | -4.082  | 5.71e-05 *** |
| Science_Fiction | -0.249838 | 0.201740 | -1.238  | 0.216523   |
| Factual_Book | 0.603207  | 0.320628  | 1.881   | 0.060889 . |
| Game         | 0.761276  | 0.519441  | 1.466   | 0.143806   |
| Religious_Texts | 0.924670  | 0.802693  | 1.152   | 0.250249   |
| Original_Screenplay | 0.220309  | 0.145011  | 1.519   | 0.129744   |
| Spin_Off     | 0.706098  | 0.673289  | 1.049   | 0.295141   |
| Live_Action  | 0.582439  | 0.295520  | 1.971   | 0.049648 *  |
| IMDb_Score   | 0.016808  | 0.010440  | 1.610   | 0.108460   |
| Rotten_Tomatoes_Score | 0.015138 | 0.003441 | 4.399   | 1.51e-05 *** |
| Sequel       | 0.659162  | 0.163031  | 4.043   | 6.70e-05 *** |</p>
<table>
<thead>
<tr>
<th>Regressors</th>
<th>Mean sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PG</td>
<td>13.220</td>
<td>10.7180</td>
<td>0.0011842 **</td>
</tr>
<tr>
<td>PG-13</td>
<td>6.529</td>
<td>5.2934</td>
<td>0.0220878 *</td>
</tr>
<tr>
<td>Adventure</td>
<td>6.909</td>
<td>5.6017</td>
<td>0.0185739 *</td>
</tr>
<tr>
<td>Drama</td>
<td>41.926</td>
<td>33.9908</td>
<td>1.422e-08 ***</td>
</tr>
<tr>
<td>Horror</td>
<td>31.156</td>
<td>25.2593</td>
<td>8.600e-07 ***</td>
</tr>
<tr>
<td>Musical</td>
<td>6.529</td>
<td>5.2931</td>
<td>0.0220907 *</td>
</tr>
<tr>
<td>C</td>
<td>13.120</td>
<td>10.6364</td>
<td>0.0012357 **</td>
</tr>
<tr>
<td>Historical_Fiction</td>
<td>13.970</td>
<td>11.3259</td>
<td>0.0008631 ***</td>
</tr>
<tr>
<td>Dramatization</td>
<td>13.701</td>
<td>11.1079</td>
<td>0.0009666 ***</td>
</tr>
<tr>
<td>Science_Fiction</td>
<td>0.038</td>
<td>0.0306</td>
<td>0.8611569</td>
</tr>
<tr>
<td>Factual_Book</td>
<td>5.796</td>
<td>4.6992</td>
<td>0.0309579 *</td>
</tr>
<tr>
<td>Game</td>
<td>0.030</td>
<td>0.0246</td>
<td>0.8755388</td>
</tr>
<tr>
<td>Religious_Text</td>
<td>0.541</td>
<td>0.4384</td>
<td>0.5084200</td>
</tr>
<tr>
<td>Original_Screenplay</td>
<td>0.196</td>
<td>0.1586</td>
<td>0.6907381</td>
</tr>
<tr>
<td>Spin_Off</td>
<td>2.574</td>
<td>2.0869</td>
<td>0.1496004</td>
</tr>
<tr>
<td>Live_Action</td>
<td>0.632</td>
<td>0.5123</td>
<td>0.4746769</td>
</tr>
<tr>
<td>IMDb_Score</td>
<td>59.549</td>
<td>48.2785</td>
<td>2.280e-11 ***</td>
</tr>
<tr>
<td>Rotten_Tomatoes_Score</td>
<td>22.776</td>
<td>18.4652</td>
<td>2.337e-05 ***</td>
</tr>
<tr>
<td>Sequel</td>
<td>20.163</td>
<td>16.3471</td>
<td>6.700e-05 ***</td>
</tr>
</tbody>
</table>

Where the significance codes are: 0 (***) 0.001 (**), 0.01 (*), 0.05 (.)
Table 5: Confidence intervals for coefficients

<table>
<thead>
<tr>
<th>coefficient</th>
<th>lower bound(2.5%)</th>
<th>upper bound(97.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-2.9639305912</td>
<td>-0.43678401</td>
</tr>
<tr>
<td>PG</td>
<td>0.4610930168</td>
<td>1.54485677</td>
</tr>
<tr>
<td>PG-13</td>
<td>0.0882485782</td>
<td>0.65128878</td>
</tr>
<tr>
<td>Adventure</td>
<td>-0.8377297147</td>
<td>0.13868218</td>
</tr>
<tr>
<td>Drama</td>
<td>-0.6991446257</td>
<td>0.06512582</td>
</tr>
<tr>
<td>Horror</td>
<td>0.8285440264</td>
<td>1.99514843</td>
</tr>
<tr>
<td>Musical</td>
<td>0.2744397374</td>
<td>2.94266953</td>
</tr>
<tr>
<td>C</td>
<td>-0.1435288399</td>
<td>0.64526127</td>
</tr>
<tr>
<td>Historical_Fiction</td>
<td>-1.3407431183</td>
<td>-0.46102547</td>
</tr>
<tr>
<td>Dramatization</td>
<td>-1.5070023679</td>
<td>-0.52669599</td>
</tr>
<tr>
<td>Science_Fiction</td>
<td>-0.6468323836</td>
<td>0.14715592</td>
</tr>
<tr>
<td>Factual_Book</td>
<td>-0.0277410610</td>
<td>1.23415486</td>
</tr>
<tr>
<td>Game</td>
<td>-0.2609054098</td>
<td>1.78345715</td>
</tr>
<tr>
<td>Religious_Text</td>
<td>-0.6549103846</td>
<td>2.50425025</td>
</tr>
<tr>
<td>Original_Screenplay</td>
<td>-0.0650512150</td>
<td>0.50566854</td>
</tr>
<tr>
<td>Spin_Off</td>
<td>-0.6188337745</td>
<td>2.03102971</td>
</tr>
<tr>
<td>Live_Action</td>
<td>0.0009006845</td>
<td>1.16397756</td>
</tr>
<tr>
<td>IMDb_Score</td>
<td>-0.0037367989</td>
<td>0.03735211</td>
</tr>
<tr>
<td>Rotten_Tomatoes_Score</td>
<td>0.0083668594</td>
<td>0.02190927</td>
</tr>
<tr>
<td>Sequel</td>
<td>0.3383404801</td>
<td>0.97998312</td>
</tr>
</tbody>
</table>

4.1.1 Residuals plotted against the fitted values

Figure 7: The residuals plotted against the fitted line for the final model
The problematic funnel shape that existed in the initial model has been handled through the Box-Cox transformation and assumption of constant variance now holds.

### 4.1.2 Normal Q-Q plot

![Normal Q-Q plot](image)

Figure 8: A Q-Q plot for the final model

The normal Q-Q plot for the final model has been improved from the initial model where the noticeable departure from the line that existed has been eliminated. The residuals are therefore normally distributed with mean zero and the final model follows the normality assumption.

### 4.1.3 Histogram of residuals

![Histogram of residuals](image)

Figure 9: A histogram of the residual for the final model
The Histogram for the final model is also an improvement from the initial model. The skewed histogram of the initial model has been corrected and the residuals of the final model are therefore normally distributed.

4.1.4 Scale-location plot

The Scale-location plot for the final model is an improvement compared to the initial model. The line has become more horizontal and the data points are more randomly equally spread which indicates that the residuals are equally spread along the range of the regressors.

Figure 10: The Scale-Location plot for the final model

The Scale-location plot for the final model is an improvement compared to the initial model. The line has become more horizontal and the data points are more randomly equally spread which indicates that the residuals are equally spread along the range of the regressors.
4.1.5 Residual vs leverage plot

Figure 11: A plot of the residual vs leverage for the final model

Since there did not exist any problems concerning this plot for the initial model it still holds for the final model.
4.1.6 Multicollinearity

Table 6: VIF for final the model

<table>
<thead>
<tr>
<th>Regressor</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>PG</td>
<td>2.596488</td>
</tr>
<tr>
<td>PG-13</td>
<td>1.303494</td>
</tr>
<tr>
<td>Adventure</td>
<td>2.588754</td>
</tr>
<tr>
<td>Drama</td>
<td>1.709011</td>
</tr>
<tr>
<td>Horror</td>
<td>1.083096</td>
</tr>
<tr>
<td>Musical</td>
<td>1.107482</td>
</tr>
<tr>
<td>C</td>
<td>1.116226</td>
</tr>
<tr>
<td>Historical_Fiction</td>
<td>1.263615</td>
</tr>
<tr>
<td>Dramatization</td>
<td>2.018982</td>
</tr>
<tr>
<td>Science_Fiction</td>
<td>1.180583</td>
</tr>
<tr>
<td>Factual_Book</td>
<td>1.416353</td>
</tr>
<tr>
<td>Game</td>
<td>1.076771</td>
</tr>
<tr>
<td>Religious_Text</td>
<td>1.038249</td>
</tr>
<tr>
<td>Original_Screenplay</td>
<td>1.368096</td>
</tr>
<tr>
<td>Spin_Off</td>
<td>1.092288</td>
</tr>
<tr>
<td>Live_Action</td>
<td>2.941398</td>
</tr>
<tr>
<td>IMDb_Score</td>
<td>2.571515</td>
</tr>
<tr>
<td>Rotten_Tomatoes_Score</td>
<td>2.443601</td>
</tr>
<tr>
<td>Sequel</td>
<td>1.204584</td>
</tr>
</tbody>
</table>

All VIF values are below 5 in the final model which means that multicollinearity will not interfere with the regression coefficients.

5 Discussion

5.1 Evaluation of final model

The first thing to note is the $R^2$ and adjusted $R^2$ values. Since the $R^2$ is equal to 0.41 the final model can’t even explain 50% of the variability in the data. This could be an indication that the regression analysis result provides a base for inference rather than prediction. By this we mean that the interpretation of significance is still relevant, i.e. how for instance a genre on average behaves, but that trying to predict profitability is futile since the amount of variance left in the data seriously hampers the interpretation of the prediction results in new observations. In this case, it is likely that the present degree of variability explanation will hold or be lower when translating to new data for prediction.

Secondly looking at the cross-validation results showed that there was a performance gap between the two variable selection approaches. This isn’t surprising since the second approach allows for greater model flexibility compared to the first approach. Among the models generated by the second approach, there wasn’t a large difference performance wise. The final model had an MSE equal
to 1.4055 which gives a RMSE equal to 1.1855. This a measure of the prediction error and it is large considering that the response represents a profitability. As an illustration, a movie that in reality generates a total box office equal to the budget, which means a response equal to zero, could be predicted to make a substantial return, basing a decision on the prediction could be catastrophic.

Given these points, the predictive power is deemed to be low. By low, we mean that the model in most cases isn’t usable when specifically making predictions for individual movies.

5.2 Impact of the regressor variables

Even though the predictive power for the final model is low the interpretation of significance is still the same. That is to say, the significant coefficients highlight the mean change in the response for one unit change in the regressor while keeping all others fixed. Important to note is also that for categorical variables the significant coefficients for dummy variables still signify the mean change in response but it is relative to the reference. In other words, it is the mean change in the response when the movie goes from belonging to the reference to the significant category.

Also important to note is that the current response is a box-cox transformation of the aggregated box office divided by the budget. Comparing different responses still has the same interpretation, i.e. if a movie A has a response larger than movie B it still means that A had a greater return compared to B’s return. This holds because the transformation function is monotonically increasing. On the other hand, the transformation of the response complicates the interpretation of coefficients effects. For instance, if a certain movie belonging to the reference category in Rating made a 0% return then the coefficient for the Rating PG then indicates how much higher the return will be. But how much higher the return will be is hard to see since the response is transformed. The coefficient for PG is 1.00 which would mean that in the case mentioned above the movie would instead have a return equal to 137%. If the coefficient instead was equal to 1.5 the movie would have made a 235% return, 0.5 gives a return of 58.6%. If the coefficient was equal to -1 the movie would instead make a -70.6% return. In addition, the interpretation for coefficients for other variables is the same, i.e. a coefficient above 0.5 means a significant increase in profit compared to if the movie made a return equal to its production budget while it belonged to the relevant reference category. The magnitude of coefficients will mainly be analyzed from this perspective.

5.2.1 Rating

Both PG and PG-13 have a positive impact on profitability although PG has the greater impact with a coefficient equal to 1.00 which means that it impacts the profitability greatly based on the perspective outlined at the beginning of 5.2. The PG-13 also impacts the profitability but not as much. A probable explanation for this is that the less restricted rating a movie gets the larger the potential audience gets. A larger potential audience would mean that more customers could see the movie which is positive. Since a lower MPAA-rating has a positive impact on the profitability it also possibly explains why production companies are striving for these ratings.
5.2.2 Genre

Within the genres of the final model Horror(1.41) and Musical(1.61) have a positive influence on profitability which means that they heavily impact profitability based on the perspective at the beginning of 5.2 while Drama(-0.32) and Adventure(-0.35) have negative ones. Horror(3e-06 ***) has the lowest P-value followed by Musical(0.02 *) which still was relatively low compared to drama(0.10) and adventure(0.16) which have higher P-values.

An explanation as to why Horror has positive influence is that horror movies are relatively cheap to make. A large portion of the movies in our dataset with a relatively small budget are horror movies. Since big name actors, expensive locations or special effects are not as prevalent in horror movies the costs can be held low. The adrenaline thrill that people gets in horror movies seems to be cheap to produce but still effective enough to get customers to the cinema. The horror genre, therefore, seems promising since it is more profitable compared to most genres but has lower risk since horror movies generally have lower production budgets.

The genre Musical has a higher influence on profitability than Horror but also a higher P-value. Something that is worth mentioning is that only three movies in the dataset were musicals and two of these movies have some of the largest influence of all the observations. This can be observed in the residual vs leverage plot in section (5.1.5) where the movies La La Land and The Greatest Showman are indicated to be outliers since they are close to the dashed lines. According to the model Musical seems to be a profitable genre but the aforementioned characteristics of this genre need to be taken into consideration while evaluating the impact the musical genre has on profitability.

5.2.3 Sequel

The regressor variable sequel which has a high level of significance(7e-05 ***) has a positive impact(0.66) on the response variable profitability. Linking back to the discussion at the beginning of 5.2 we see that sequel also has a significant impact on profitability. This result backs up the claim of increasing amounts of sequels which was stated in the introduction. If being a sequel to an already established movie has a positive impact on profitability it makes sense for the production companies to continue creating sequels.

The explanation of this positive impact is that sequels already have an established audience which means that it is not as important to inspire interest into new customers. The risk of failure is lower when creating sequels since if the previous movies were successful the majority of the established audience are likely to watch the new movie.

5.2.4 RottenTomatoesScore

This regressor had a high significance(1.5e-05 *** in the final) both in the initial model and in the final model. Since the regressor isn’t categorical but continuous the interpretation of it is different. The coefficient(0.0151) of the regressor is positive meaning that an increase in score leads to an increase in profitability. A 100% score means almost 1.5 increase in the response and this indicates a great increase in profitability.

The result is especially interesting since it provides a clear proof of a link between critical suc-
cess and business-centric success. Important to realize is that this in some sense also links box office success, a measure of public reception, and critical reception.

5.2.5 IMDbScore

The IMDb Score seems to have the same impact (0.0168) as the Rotten tomatoes score mentioned above but has a higher P-value (0.11) and is therefore not as statistically significant.

The difference between the two types of scores is that the IMDb Score is based on regular peoples reviews while the Rotten tomatoes score is based on film critics reviews. This difference could possibly explain the difference in the significance of the two regressor variables. It could be that the Rotten Tomatoes Score is a more used score for determining to see a movie in the cinema or not. Since the Rotten tomatoes score is based on film critics reviews it means that most of the reviews that the score is based on will be from the time period of the movies release date. The reviews that the IMDb score is based on is on the other hand made by the average movie-goer and can be made at any time, even after the movie is not being played at the cinemas anymore.

This difference could affect the significance of the scores in different ways. The different nature of the scores could affect which type of score people rely on when deciding which movie’s to see. It could also be that the rotten tomatoes score of movies has remained relatively stable since the period of when it was played in cinemas while the IMDb score has changed over time. Since all the scores for the movies were collected in 2018 this could mean that the collected IMDb scores do not reflect the scores the movies had at release which could explain the difference in significance.

5.2.6 DirectorScore

Only the category C of DirectorScore was included in the final model and it was not significant (0.21). This result is interesting since it indicates that there is no link between the director and profitability or rather that there is no link between our measure of the director and profitability. It seems reasonable that the director affects profitability in some way, a new measure of the director may need to be developed.

5.2.7 CreativeType

All of the different creative types included in the final model have a negative impact on profitability. It is also mentionable that both Historical Fiction (7.e-05 *** ) and Dramatization (6e-05 *** ) have low P-values while Science Fiction (0.22) has a higher P-value.

Both Historical Fiction and Dramatization have negative coefficients indicating that these types of movies perform worse compared to those in the reference category. Since both coefficients are close to -1 their negative impact on the profitability is great. Considering the creative types of modern blockbusters such as Star Wars VII or the Marvel movies, and the fact that super_hero and science_fiction is included in the reference then this result is not surprising.

5.2.8 Source

All the types of sources have a positive impact on profitability where religious texts (0.92) has the largest and original screenplay (0.22) has the smallest. The majority of the sources have high P-
values where the source *Factual Book* (0.06) has the lowest P-value and is quite a bit lower than the rest.

Movies based on religious texts, games, factual books or spinoffs are of similar nature, where they all have an established audience, and all of them have a greater impact than an original screenplay. This could further support the claim that was stated in the introduction, and somewhat discussed in section (6.2.3), where movie adaptations from already established media as books and games have become more common.

The positive impacts explanation is the same as the one presented in section (6.2.3) where an already established audience is an easy way to make sure that a movie is going to be profitable. When adapting an already produced and consumed story it is not just the audience that is established but the specific story in question has been tested beforehand.

### 5.2.9 ProductionMethod

The category from ProductionMethod included in the model was *Live Action* which also was significant (0.05*). The coefficient (0.58) was positive meaning that *Live Action* performs better than those in the reference category and affects profitability significantly. Most movies made today belong to the *Live Action* category, 273 movies belonged to this category in the dataset used.

Once again discussing established audiences shows that the established audience in some sense is larger for *Live Action* since most moviegoers are used to seeing this kind of movie compared to some of the niche categories in the reference such as *StopMotion Animation*. On the other hand, *Digital animation* is included in the reference which means that movies such as *Inside Out*, *Moana* and *Finding Dory* are included in the reference. Even though these mainstream animated movies perform well the *Live Action* category may outperform the reference due to other niche categories and the sheer size of *Live Action* audience. Even if *Live Action* outperforms the reference category it does not imply that the reference category is not profitable.

### 5.2.10 Confidence intervals of the regressors

Examining the 95% confidence intervals shown in table 5 for the coefficients supports the discussion above. As noted a coefficient above 0.5 indicates a major positive change in the response and in turn the predicted revenue. Similarly, a coefficient below -0.5 indicates a major negative change in the response. So if the confidence interval for a coefficient contains values close to zero it indicates that the coefficients effect on the response is questionable since it has a small effect on the response and in turn the profitability.

PG has a lower bound close to 0.5 indicating a major positive effect on the response. PG-13 has a lower bound close 0.088. If the coefficient is equal to 0.088 the effect on the response will not be relevant. The confidence interval for PG-13 shows some ambiguity regarding its effect on the response. Horror has a lower bound above 0.8 indicating that it has a major positive effect on the response. Musical has a lower bound barely above 0.27 which indicates that it has a relevant positive effect on the response. Historical Fiction has an upper bound below -0.45 indicating that it has a major negative effect on the response. Dramatization has an upper bound below -0.5 which also indicates that it has a major negative effect on the response. *Live Action* has a lower bound
above 0.0009 indicating that its effect on the response is questionable. RottenTomatoes_Score has a lower bound above 0.008. With this in mind and noting that this variable is continuous in the range of 0 to 100, it is clear that the variable still has a major effect on the response. For instance, if the coefficient was equal to the lower bound then a movie with a RottenTomatoes_Score equal to 100% would have the same effect as a dummy variable with a 0.8 coefficient. Sequel has a lower bound above 0.3 indicating that it has a relevant positive effect on the response.

5.3 Limitations of model and approach

5.3.1 Budget estimations

To examine profitability it is necessary to have revenue and cost. For movies, revenue is easy to find since box office numbers are widely available. Interpreting these box office values is also straightforward. Cost, on the other hand, is much harder to examine for movies since that budget information in most cases isn’t publicly available. In the cases that it is publicly available the interpretation of it can be problematic due to Hollywood accounting where costs can be inflated or reduced due to various reasons.

The budgets in this thesis all come from Opus Data which as far as we are concerned provide the most accurate budget estimations. It would be of interest to examine how these budgets are estimated and how reliable these estimations are. Creating estimations was beyond the scope of this thesis.

5.3.2 Marketing

An important aspect of movie creation is the marketing or rather how much is spent on the marketing. An approach to this problem is to specify that a movie only is profitable once it generates revenue larger than the production budget plus some percentage of the production budget. Actually assessing if a movie is profitable or not based on the results of this thesis becomes difficult since where to place the profitability cut-off has not been analyzed. We are as of now not aware of any sources with reliable information regarding marketing budgets. Adding to the problem of finding a cut-off is among other things fees from distributors and cinemas that for instance also could take a share of the box office revenue. This thesis consequently focused on relative profitability changes.

Aside from interpreting profitability marketing budget and marketing itself could also provide valuable information as regressors. Knowing how many watch the final trailer for a movie could, for instance, be a relevant measure.

5.3.3 Casting

This thesis did not reflect what type of casting a movie had. It is not unreasonable to think that for instance, a famous actor cast in a movie could increase its draw and its profitability. Developing some kind of regressor reflecting the casting is definitely something that would improve upon this work. There exists a lot of information regarding celebrities as opposed to the marketing budget dilemma. The problem is rather figuring out what to capture, for instance, how profitable an actor has been or how well known they are in the average household.
5.3.4 Initial model formulation

The initial model formulation in this thesis tries to predict the specific profitability of a movie. Consequently, the low predictive power could be due to the fact that the model is trying to exactly predict profitability.

Shifting focus to predicting if a movie is profitable or not with could be a better approach. Logistic regression with a binary variable representing if the profit is above some cut-off could be used to generate probabilities for movies being profitable. On the other hand, this approach limits the amount of information that the model attempt to deliver and puts a greater emphasis on deciding what constitutes a profitable movie. At the same time if a model fails to deliver reliable predictions it would be preferable to use a simpler but more reliable model. Where a movie’s profitability is in relation to the profitability cut-off is the most important piece of information in the end for making investment decisions.

5.3.5 Limitation due to scope

The regression analysis performed looked only at movies produced in the time span 2015-2017. It is possible that the results achieved mainly depended on the choice of time span and that those significant regressors would change if one tried to use future or past observations. Examining how the regression analysis results change with the time-span would be of interest since it could shed some light on the effect of trends and if there exist any truths in the movie industry resistant to the passing of time.

5.4 Conclusion

The final model managed to explain 41% of the variability in the data and achieve an MSE equal to 1.4055 which led to the conclusion that the model was not sufficient to use for future predictions. On the other hand, the final model included several significant coefficients which allowed inference. The following coefficients were significant on at least a 5% level: Horror, Musical, PG, PG – 13, HistoricalFiction, Dramatization, RottenScore and sequel. Musical had the largest positive influence on the response but contained few observations which led to uncertainty regarding how this result would translate to other musicals. A movie with a Tomatometer score of 100% had the second highest positive influence on the response. This result was especially interesting since it linked critical reception and business-centric success. HistoricalFiction and Dramatization had the most negative impact on the response which likely depended on that these kinds of movies seldom become mainstream blockbusters. Notable variables not found to be significant were among others DirectorScore and IMDbScore. In contrast, there were some limitations with this thesis which mainly were choice of time span, budget estimation accuracy, omitted regressors and model specification. Despite the issue of budget estimation accuracy, we feel that further research is warranted. Examining specifically how and which results could be leveraged for investment decisions is something that we also believe should be investigated further.
6  Porter’s five forces analysis

6.1  Introduction

As specified in the first chapter introduction the final step of movie production is distribution. For big-budget movies, the distribution company is usually under the same corporate umbrella as the production company. An example of this is the Walt Disney Studios Motion Pictures which is an American distribution company owned by The Walt Disney Company.\(^{28}\) On the other hand, the production company can be different from the distribution company. Important profitability influencing decisions in this phase are among other marketing and how the movie will be made available for viewing. Distribution companies also decide how a movie will be distributed when the theatrical run ends.

To understand what influences profitability from the distribution company perspective it is important to examine the relationship between the distribution company and the movie theatres. On one hand, the movie theatres are a customer for the distribution company since they can buy the rights to show the movie for a certain time. At the same time, movie theatres can be seen more as a partner or distribution channel since the distribution company and the movie theatre can agree to split the revenue among them. An example of this is Star Wars: The Last Jedi where movie theatres had to give Disney 65% of the ticket revenues.\(^{29}\) In the light of this and the fact that cinema is the main avenue for almost all initial mainstream movie releases the profitability from the distribution companies perspective is influenced by the agreement with the movie theatres.

This thesis has up until this point focused on how the intrinsic values of a movie affect its profitability in a cinematic release. In other words, no aspects surrounding a movie, such as different types of distribution, has been taken into consideration. The results achieved so far are mainly of interest for investment decisions during the development, pre-production and production phase but there are still important decisions to make during the distribution phase.

As said most mainstream movies as of today have an initial theatrical release. This will probably not change drastically in the coming years but it could still be of interest for distribution companies to gain knowledge regarding how the movie theater industry might change and how it could affect the distribution companies’ and their future theatrical releases. Similarly, information regarding potential substitutes for a theatrical release could be of interest.

6.2  Scope

The aim in this part of the thesis is to perform an exploratory and qualitative analysis and identify relevant actors in the movie theatre industry and discuss their possible effects on profitability for distribution companies. Consequently, no quantitative analysis or interviews will be performed. The thesis put the greatest emphasis on the regression analysis which in turn limited the potential scope, in terms of how thoroughly the problem statement is scrutinized, of this second part. This


means that the Porter’s five forces analysis mainly acts as a first step to answer and analyze the problem statement.

The analysis will be limited to the North American domestic market. This is done to simplify the analysis process. For one thing, analyzing several countries means different major movie theatre companies, different rating systems, varying movie habits and more. On the other hand, the domestic market is quite large which means that a significant part of the movie industry is still researched. For instance, the top 5 movies based on all time box office all had at least 25% coming from the domestic market. In like manner Star Wars: The Force Awakens, number 3 on all time box office, even had 45.3% coming from the domestic market.

It also important to note that this thesis will not perform the Porter analysis from the perspective of the distribution company even if this is the main party of interest. The reason for this is that the theatre industry and the distribution companies are tightly interlinked and increasing the distribution companies understanding of the theatre industry could provide valuable information for investment decisions and negotiation with the theatres. For instance, a substitute for theatres could be a potential buyer for a distribution company.

6.3 Problem statement

- Is a theatrical release the most profitable when a movie is initially released, if not what kind of distribution substitutes exist?

6.4 Method

Porter’s five forces is a relatively subjective tool but the analysis will, as long as it is possible, be based in qualitative and quantitative secondary data. Consequently, some parts of the analysis will be based on our own reasoning and will thus be somewhat subjective. The process of gathering information regarding the movie theatre industry was done relatively ad hoc although data and articles from trusted institutes and organizations was of main concern.

6.4.1 Types of movies

This part of the thesis is mainly interested in two types of movies. On one hand movies with budgets exceeding $80 million, significant marketing budgets, major studio backing, and large audiences. Example of this kind of movie is Star Wars The Last Jedi, Wonder Woman, Moana, Inside Out and The Lego Batman. On the other hand movies with budgets up to $50 million that can be considered as niche, experimental or innovative. Examples of this kind of movie are Get Out, Annihilation, Split and Lady Bird. Important to note is that this type of movie still can generate a lot of revenue, for instance, Get Out had a $255 million total box office. The former is referred to as mainstream movies and the latter as low profile movies. Both types of movies usually have theatrical releases but differ in terms of marketing, size of production and distribution company and more.

30Box Office Mojo. All Time Box Office. http://www.boxofficemojo.com/alltime/world/
31ibid
6.4.2 Choice of Porter’s five forces analysis

Porter’s five forces analysis was specifically deemed to be suitable since it is a tried and tested and easy to use framework which is suited for our problem statement. As mentioned in section 6.2, the goal was not to perform a quantititative analysis or similar which means that Porter’s five forces analysis is suitable with regard to scope and depth of analysis. To conclude, the goal is to use a tool to analyze the movie theatre industry and in turn relate the analysis to profitability for a distribution company. Since Porter’s five forces can be said to rate the attractiveness of an industry by analyzing stability, rivalry, substitutes and more it becomes evident that it is sufficient to use a tool for analyzing the movie theatre industry.

6.5 Theory

*Porter’s five forces*, developed by Michael Porter, is a method for analyzing the competition within an industry. According to Porter there exist five forces that affect company’s profitability within an industry. The five forces are the following.

- **Threat of new entrants**
  If a business is profitable it will always attract new companies which pose a threat since new entrants will grab market shares and decrease the profitability of the business. The threat of new entrant depends on existing entry barriers such as expensive infrastructure requirements or scale advantages within established companies. If the barriers are high, then the threat of new entrants is considered to be low.

- **Threat of substitutes**
  Substitute products can limit the potential of an industry by placing an upper limit on the price of the product, if the price of the product is above that of the substitute customers could choose to just buy the substitute. This would mean less growth and profit in the industry.

- **Bargaining power of customers**
  Customers with a lot of power are able to influence the price of the product negatively or demand better quality or service. This would probably negatively influence the industry profits. Powerful customers are among other aspects characterized by large volume purchases, mainly interested in standardized products that many suppliers can deliver and that the purchased product constitutes a majority of the total cost for the buyer.

- **Bargaining power of suppliers**
  The bargaining of the suppliers can be a threat since they are in control of supplied goods where they can raise the price for example. The bargaining power of the suppliers can depend on factors such as how many suppliers exist or the uniqueness of the product.

- **Industry rivalry**
  Companies within an industry can compete in many ways which will decrease the profitability. They can change prices, offer products of different quality or use other tactics to attract customers. Factors that determine the industry rivalry can be the number of competitors, market growth, and exit barriers.

---

6.6 Five forces applied on the cinematic theater industry

6.6.1 Threat of new entrants

A measure of size for the movie theatre industry is number of screens, the total number of screens in the U.S. was 40,426 in 2017. Noting that the number of movie screens in Canada 2016 was 2600 gives the number of movie screens in North America approximately equal to 43,000 (43,000 is the value the thesis will use from now on). On the other hand, the largest movie theatre chains with respect to the number of screens are AMC theatres with 8,218 screens, Regal Entertainment Group with 7,379 screens, Cinemark USA, inc. with 4,544 screens and Cineplex Entertainment LP with 1,683 screens. Consequently, the top 4 movie theatre chains constitute 50.8% of the North American market by number of screens.

Actual barriers inhibiting new entrants are among other capital requirements. Movie theatres require capital upfront to finance the purchase of real estate and cinema technology. Fixed costs can also be high due to licensing fees for movies and potential rent. Important to note is also that to compete with the 4 major chains several screens have to be financed, for instance, the capital requirements for 1000 screens are likely very high. Capital requirements indicate a high barrier to entry.

Another threat to new entrants is economies of scale but how this translates to movie theatres is arguable. Increasing the number of screens should lead to a linear increase in costs since real estate and cinema technology still has to be bought. On the other hand, major movie theatre chains are purchasing larger volumes from their suppliers which reasonably means better business agreements. As an illustration, AMC are likely to negotiate better deals with Dolby Cinema which provides high-quality cinema experiences since AMC owns such a large market fraction. The size advantage probably translates to more secure and better distribution channels. Economies of scale indicate medium barrier to entry.

In spite of the mentioned threats differentiation can offer new entrants some possibilities. New entrants can target niche markets such as luxury cinema, specialized settings or events since these can be less saturated. Still important to note is that this is most probable only on a small scale. Differentiation can also be detrimental to new entrants if the major chains have a strong market recognition. This scenario is best illustrated with Coca Cola where a new entrance in the soft drinks industry will suffer due to them not being Coca Cola. On the other hand, it can be hard to judge this effect in the movie theatre industry. The four major chains are probably the most recognizable due to their number of screens but differentiation is low in some sense since the chains show similar movies. It is arguable whether or not differentiation offers any possibilities for new entrants apart from small scale ones.

There can also be advantages independent of size for established chains that new entrants can’t

---


utilize. An example of this is favorable locations. Established chains are likely to have secured the best locations for the cinemas which prevent new entrants from utilizing the best locations. This indicates a higher barrier to entry.

To conclude we see that economies of scale, capital requirements and advantages independent of size indicates high barriers of entry for new large scale entrants. In the light of the fact that 4 major chains own approximately 50% of the total number of screens, the barrier of entry is deemed to be high.

### 6.6.2 Threat of substitutes

The movie theater business has changed a lot since its infancy. The percentage of the population that goes to the movies weekly has changed from 65% in 1930 to 10% in 1964 and has stayed around 10% since.

A reasonable explanation for this is the growth of other types of entertainment that can be enjoyed in the home like TV. It seems like the movie theater business has been affected by a substitute before and the threat of substitutes still exists today.

Streaming services like Netflix could pose a threat to the movie theater industry since it provides the same movies, although later than the movie theaters and even Netflix’s own exclusive made movies. The newer generations seem to adopt the streaming way of consuming entertainment where 67% of millennials pay for streaming services but the adoption of streaming services does not seem to affect the millennials movie theater attendance. This indicates that the movie theaters provide a different kind of experience and that streaming services is not a big threat. Harking back to the stability of the percentage of people going to the movies weekly it seems clear that the movie theatre industry already has lost its share of customers that were seeking a different kind of experience. It is likely that streaming services instead pose a threat to TV.

The cinematic experience is the more immersive way to watch a movie with the large high-quality screen and sound systems and it seems people still want to see a movie in the best way possible. This experience can be emulated at home with home cinema systems. Even so, the portion of people that choose home cinema systems instead of movie theatres is likely small due issues of cost and technological investment.

Movies going straight-to-video has for a time now been an avenue for distributing movies and can be seen as a substitute for theatrical releases. But this was often the case for independent filmmakers, movies directed at niche segments or movies that lacked support from studios. As such this avenue is associated with movies of a lesser quality compared to most with a theatrical release. Although this may be true movies today going straight-to-streaming might not be associated with the same kind of stigma. An example of this is the movie The Cloverfield Paradox which

---


38 Meaning movies released directly to video instead of having a theatrical release

39 Meaning movies directly released on streaming services instead of having a theatrical release
aired a surprise trailer during Super Bowl LII. The movie was released on Netflix shortly after the game. This was an extraordinary marketing stunt by Netflix since the Super Bowl trailer meant that they reached one of the biggest TV audiences and at the same time managed to deliver the movie the same day. On the other hand, the movie was originally planned to have a theatrical release by paramount pictures but paramount pictures decided to drop the movie. This might indicate that the movie was anticipated to have a poor theatrical release. Another example is the movie Annihilation which had a theatrical release in North America done by Paramount pictures and released digitally on Netflix in the rest of the world. The movie reportedly didn’t test well but the director refused to make any changes. Paramount Pictures might have turned to streaming in the hopes that it could increase the potential of the movie. So movies going straight-to-streaming might not be a strong substitute for movies with large theatrical releases but rather a new avenue for distribution lying somewhere between straight-to-video and a theatrical release.

6.6.3 Bargaining power of customers

The product that is consumed is the same for all the different movie theater companies since they screen the same movies. Some differentiation can exist with different levels of technologies like IMAX and sound systems or different levels of luxury like more luxurious seats but in the end it’s still the same movie. Promotion is also an important way to differentiate. One of the most influential factors when choosing movie theaters is probably the geographical location. So the product is standardized and consumers could get the product from other theaters but this is not such a major problem since theaters are generally widely spread.

The movie theaters customers will generally watch relatively few movies per year and since the movie theater industry is a business to consumer business it does not have the kind of large volume customers that exist in other industries. Individual customers will, therefore, not have great bargaining power.

6.6.4 Bargaining power of suppliers

The main suppliers for the movie theaters are the film studios that provide the movies. Movies from large production companies, mainstream movies, are essential for the movie theaters and the bargaining power of the suppliers is therefore strong. Since the movie theaters want to screen the big movies which attract viewers the production companies can usually pressure them when making the deal of how the ticket sales are supposed to be split up. An example of this is Disney who took 65% of ticket sales for Star Wars: The Last Jedi and made theaters screen the movie for at least 4 weeks. The bargaining power increases with the size of the production company and also increases with smaller movie theater companies.

---


Another problem is that suppliers can’t be changed in the same sense as with other industries with several suppliers of the same items. The movie theaters want to screen all the popular movies and it is, therefore, harder for the movie theaters to handle the bargaining power of the suppliers. Every movie is an unique product that can only be supplied by one supplier and this further increases the severity of the bargaining power of the suppliers. Since the movie theaters merely screen already finished movies they are dependent on the suppliers who determine the quality of the movies.

But at the same time, a distribution company trying to sell a low profile movie to a movie theatre chain might not have the same type of bargaining power. If the movie theatre has low expectations for the low-profile movie then the movie theatre could turn down an offer from the distribution company since the low-profile movie probably won’t generate many ticket sales. The bargaining power shifts when the expectation of ticket sales declines.

### 6.6.5 Industry rivalry

Theaters screen the same movies and the product, therefore, lacks differentiation which could imply that there exists high rivalry since the customers can switch between the theaters and consume the same product. The main influencing factor for choosing a theatre is probably the geographical location as mentioned in section (6.4.3). This means that geographical location is an area where the movie theater companies can compete within to attract the most customers. The lack of differentiation could also put a greater emphasis on promotion and marketing.

Interesting to note is that the movie industry as of today mainly employs a uniform pricing structure meaning that different movies will have the same ticket price. They can vary however due to things such as 3D. None the less the uniform pricing structure in some sense limits the opportunities for movie theatre chains to use innovative pricing strategies. Price competition is probably not as prevalent. Distribution fees and agreements could also limit the potential of price competition.

Slow industry growth is also something that can contribute to an increased rivalry. Tickets sales have remained approximately the same from 2011 to 2017\(^44\). In like manner, the domestic total box office had an annual growth of 0.55% from 2010 to 2017\(^45\). Comparing this to the projected total growth of $38 billion in 2016 to $50 billion in 2020 mentioned at the beginning of this article shows that the domestic market experiences slow growth which increases the industry rivalry\(^46\). Although, the industry rivalry is in the end deemed to be low.

### 6.7 Discussion

An answer to the problem statement has been found through the analysis of Porter’s five forces applied on the movie theater business. The threat of new entrants is considered to be low which favors the movie theaters. The low threat could be positive for the movie studios since their distribution channel is stable and would suggest movie theaters as a good way to distribute from

\(^{44}\text{The Numbers. Domestic Movie Theatrical Market Summary 1995 to 2018}} \quad \text{https://www.the-numbers.com/market/} \\
^{45}\text{Box Office Mojo. Yearly Box Office.}} \quad \text{http://www.boxofficemojo.com/yearly/?view2=domestic&view=releasedate&p=.htm} \\
this aspect. Film studios could on the other hand benefit from a high threat of new entrants since it would mean that new smaller movie theater chains enter the market. Since these theater chains would be smaller than the existing chains the studios would have greater bargaining power over them and would be able to pressure them and negotiate more profitable deals. So the film studios could benefit from a movie theater business with smaller but more numerous theater chains.

The conclusion regarding possible substitutes like streaming services or straight-to-video movies is that they cannot replace the movie theaters since it offers a different kind of experience. Streaming services that usually stream movies that already have been shown in the movie theaters does not pose a threat. Streaming services have also started to show their own exclusive big-budget movies on the same level of quality of the ones shown in the theaters. This is a possible substitution for the theaters. It also creates a different situation for the film studios since the streaming services already have paid for their exclusive movies before streaming them. Since the film studio and streaming services can’t split the ticket sales like with movie theaters the film studios probably get an upfront payment from the streaming services. This means that film studios are guaranteed a profit regardless of how the movie is performing.

The bargaining power of the suppliers is considered to be high, especially towards smaller theater chains. This makes movie theaters an attractive distribution channel for movies since the film studios can acquire beneficial deals. But the bargaining power also depends on the size of the movie that the distribution company is trying to sell and the timing of it. For instance, a niche movie might not be able to pressure major theatre chains into beneficial deals instead the distribution company might be the ones pressured to even achieve a theatrical release. Also, a distribution company might try to sell the rights for a movie some weeks after the initial theatrical release. The possible audience, in this case, is likely small compared to the audience during the launch which means that the theatre chain is less willing to negotiate. In this case of smaller movies or after theatrical release deals straight-to-streaming could be a potential substitute. In spite of this, it is important to realize that the concept of bigger straight to streaming movies is just emerging.

The bargaining power of the customers is considered to be low. The low risk can be attributed to the business to customer relationship where the customers are numerous and buy in small volumes. This is positive for the movie theaters and in turn the film studios since it means greater stability and further implies that theaters are a good distribution channel.

The industry rivalry is considered to be low. The low risk is positive for both the theatres and film studios in the sense that no noteworthy price competition exists which otherwise would have decreased the profitability of the business. Strategies that focus on attracting more customers and creating more demand without bringing down the prices can be positive both for the theaters and film studios. Since the theaters are limited regarding the product and pricing they have to focus on other areas as the promotion and marketing of the theaters and the quality of the theatres themselves. A rivalry that focuses on these aspects would increase the profitability of the business and is the kind of rivalry that exists between theatres chains. This would further argue for the movie theatre as a good distribution channel.
6.8 Conclusion

The analysis of all five forces leads to the conclusion that movie theaters are well suited to be a distributor for mainstream movies that come from well-known studios. This was the case since the negotiation power for suppliers of mainstream movies was deemed to be high and that the movie theatre appeared to be healthy in terms of rivalry. The high barrier of entry in the movie theatre industry also implies that suppliers of mainstream movies should focus on the relations with major movie theatre chains. No significant substitutes for the cinema experience were found.

In contrast is the situation for low profile movies. Movie theaters could be argued are not as suitable for distribution companies which intend to release low profile movies since they generally lack the negotiation power of mainstream movies. An alternative for these studios is to release their movies on streaming services directly which could prove to be more profitable than a theatrical release.

This thesis has in conclusion identified influencing factors which affects movie profitability and a regression model has been created which can be used to understand and predict movie profitability. These factors are determined by the decisions connected to the movie itself. The decision of how the movie is going to be distributed has also been analyzed through Porter’s five forces analysis on the movie theatre business. This has been done to determine how suitable movie theatres are as a distribution channel for different kind of movies. Movie theatres was deemed a satisfactory distribution channel for major movies but not as satisfactory for low profile movies.

6.9 Limitation of approach

This analysis has focused on the profitability of the initial release of a movie. Consequently, how the movie is made available for home-viewing after the theatrical release has not been discussed. DVD-sales and similar can generate a lot of revenue for the distribution companies. A theatrical release may be utilized as a marketing tool for increasing the revenue from home-viewing. This means that maximizing profitability in the initial release decreases the overall profit for the distribution company. A study done in Australia found a high correlation between theatrical box office and DVD revenue. Further research regarding if more lenient negotiations with movie theatres could increase sales from home-viewing and in turn increase overall profitability is recommended.

This kind of analysis is also mostly subjective even though articles have been used while estimating the five forces. We tried to base the analysis on facts and articles but how these are interpreted within the framework is still subjective which means that conclusions can vary from author to author. This is generally not desirable in a scientific paper and limits the reliability and validity of the second part of the thesis. Other alternative approaches exist to determine the suitability of movie theatres as a distribution channel which are not as subjective. If the necessary data can be acquired a regression analysis could have been used to investigate the problem statement. Interviews with relevant people could also be used as sources to gather more information about the business. The Porter analysis is a first step to understand the industry and further research through approaches like the ones stated above could possibly result in a different conclusions. At the same

---

time, we stress that we do not see the conclusions from the Porter analysis as an alternative to a quantitative analysis or an interview based analysis but as mentioned before a jumping off point for further research. An example of how the conclusions can be used for further research is to perform a regression analysis comparing movies with a theatrical release and a streaming release.
References


