ISQA 4010: Business Intelligence

An Analysis of Movie Data

Data Analysis Report

Miguel Rocha, Aaron Strong, Edwin Echenim, Marcus Spotanski, Naveenkumar Kantem

astrong@unomaha.edu

eechenim@unomaha.edu

mspotanski@unomaha.edu

miguelrocha@unomaha.edu

nkantem@unomaha.edu

Table of Contents

Executive Summary 3

Dataset Description 4

Figure 1.1

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	imdb_id	title	budget	revenue	certification	popularity	vote_count	vote_average	runtime	genres	release_date	release_year
2	631842	tt15679400	Knock at the Cabin	20000000	52000000	R	3975.228	866	6.476	100	Horror, Mystery, Thriller	2/1/2023	2023
3	505642	tt9114286	Black Panther: Wakanda Forever	250000000	858535561	PG-13	3095.305	3902	7.347	162	Action, Adventure, Science Fiction	11/9/2022	2022
4	315162	tt3915174	Puss in Boots: The Last Wish	90000000	442000000	PG	2589.341	4339	8.413	103	Animation, Adventure, Comedy, Family	12/7/2022	2022
5	646389	tt5884796	Plane	25000000	51000000	R	2068.255	802	6.859	107	Action, Adventure, Thriller	1/12/2023	2023
6	677179	tt11145118	Creed III	0	100000000	PG-13	1920.868	191	7.139	116	Drama, Action	3/1/2023	2023
7	1058949	tt13614388	Little Dixie	0	0	R	1688.474	56	6.402	105	Thriller, Action, Drama, Crime	2/3/2023	2023
8	937278	tt7405458	A Man Called Otto	50000000	103842303	PG-13	1545.382	522	7.9	126	Comedy, Drama	12/28/2022	2022
9	772515	tt11718516	Huesera: The Bone Woman	0	0		1436.445	84	6.631	93	Drama, Horror, Mystery	2/10/2023	2023
10	1077280	tt26198528	Die Hart	0	0		1386.983	102	6.627	85	Action, Comedy, Thriller	2/22/2023	2023
11	842942	tt9853500	Bandit	0	0	R	1379.473	98	6.306	126	Crime, Drama, Thriller	9/23/2022	2022
12	536554	tt8760708	M3GAN	12000000	167643991	PG-13	1319.223	1915	7.496	102	Science Fiction, Horror, Comedy	12/28/2022	2022
13	76600	tt1630029	Avatar: The Way of Water	460000000	2281000000		1318.499	5631	7.739	192	Science Fiction, Adventure, Action	12/14/2022	2022
14	823999	tt15307230	Diabolik - Ginko all'attacco!	0	0		1308.043	71	5.796	111	Crime, Action	11/17/2022	2022
15	758009	tt9686790	Shotgun Wedding	0	4495365	R	1207.329	607	6.362	101	Action, Romance, Comedy	12/28/2022	2022
16	640146	tt10954600	Ant-Man and the Wasp: Quantum	200000000	419000000	PG-13	1032.078	913	6.407	125	Adventure, Science Fiction, Comedy	2/15/2023	2023
17	267805	tt1655444	There Are No Saints	0	0		952.788	66	5.553	105	Action	5/27/2022	2022
18	436270	tt6443346	Black Adam	200000000	392952111		841.29	4369	7.147	125	Action, Fantasy, Science Fiction	10/19/2022	2022
19	1035806	tt22394702	Detective Knight: Independence	0	0	R	818.828	81	6.012	92	Action, Crime, Thriller	1/20/2023	2023
20	785084	tt13833688	The Whale	3000000	32000000	R	799.218	830	8	117	Drama	12/9/2022	2022
21	842544	tt14873054	Transfusion	0	0		755.138	127	6.602	105	Crime, Thriller, Drama	1/5/2023	2023
22	906221	tt16280138	Magic Mike's Last Dance	45000000	52000000	R	746.568	107	7.098	112	Comedy, Drama, Romance	2/9/2023	2023
23	840789	tt12390388	Matadero	0	0		705.46	1	3	106	Drama	12/9/2022	2022
24	1003580	tt22769820	Legion of Super-Heroes	0	0	PG-13	691.55	60	6.692	84	Animation, Action, Science Fiction	2/7/2023	2023
25	676547	tt9271672	Prey for the Devil	0	37655699	PG-13	666.726	543	7.2	93	Horror, Thriller	10/23/2022	2022
26	653851	tt7693316	Devotion	90000000	20000000	PG-13	655.399	353	7.381	139	War, History, Drama	11/23/2022	2022
27	996727	tt15441472	The Price We Pay	0	0	R	630.425	84	6.1	85	Action, Thriller, Mystery, Crime, Horror	1/13/2023	2023

..... 5

Figure 1.2

	N	O	P	Q
1	production_companies	production_companies_country	language	country
2	Blinding Edge Pictures, Universal Pictures, FilmNation Entertainment, Wishmore, Perfect World Pictures	US, US, US, US, CN	en	United States of America
3	Marvel Studios	US	en	United States of America
4	DreamWorks Animation, Universal Pictures	US, US	en	United States of America
5	MadRiver Pictures, Di Bonaventura Pictures, G-BASE, Olive Hill Media, Riverstone Pictures	US, US, US, US, GB	en	United States of America
6	Metro-Goldwyn-Mayer, Proximity Media, Balboa Productions, Outlier Society Productions, Chartoff-Winkl	US, US, US, US	en	United States of America
7	Roxwell Films, Three Point Capital	US, US	en	United States of America
8	Playtone, SF Studios, 2DUXÂ², Stage 6 Films, Artistic Films, STX Entertainment	US, SE, US, US, US	en	Sweden
9	Disruptiva Films, Machete Producciones, Maligno Gorehouse, Terminal	MX, MX, MX, MX	es	Mexico
10	HartBeat Productions	US	en	United States of America
11	Goldrush Entertainment, Yale Productions, Lucky 13 Productions	CA, US, US	en	Canada
12	Atomic Monster, Blumhouse Productions, Divide / Conquer, Universal Pictures	US, US, US, US	en	United States of America
13	20th Century Studios, Lightstorm Entertainment	US, US	en	United States of America
14	Mompracem, Astorina S.A.C., Emilia-Romagna Film Commission, RAI Cinema	IT, IT	it	Italy
15	Lionsgate, Mandeville Films, Nuyorican Productions, Maximum Effort	US, US, US, US	en	United States of America
16	Marvel Studios, Kevin Feige Productions	US, US	en	United States of America
17	Open Window Productions, BN Films, Itaca Films, Itaca Films	MX, US	en	United States of America
18	New Line Cinema, Flynn Picture Company, DC Films, Seven Bucks Productions	US, US, US, US	en	United States of America
19	Arcana Studio, 308 Enterprises, BondIt Media Capital, Buffalo 8	CA, CA, US, US	en	Canada
20	A24, Protozoa Pictures	US, US	en	United States of America
21	Deeper Water Films, Madman Entertainment, Altitude Media		en	Australia
22	Warner Bros. Pictures, Nick Wechsler Productions, Free Association	US, US, US	en	United States of America
23	El Viaje Films, 4 Ã 4 Productions, Prisma Cine, Magoya Films	ES	es	Argentina
24	Warner Bros. Animation, DC Entertainment, Warner Bros. Home Entertainment Group	US, US, US	en	United States of America
25	Lionsgate, Gold Circle Films, Confluence Productions	US, US	en	United States of America
26	Black Label Media, STX Entertainment, Columbia Pictures, Stage 6 Films	US, US, US, US	en	United States of America
27	828 Productions, Buffalo 8, 828 Media Capital, VMI Worldwide, ABS Payroll & Production Accounting Servi	US	en	United States of America

..... 5

Tools and Techniques 6

TOOLS 6

Excel 6
R Studio 6
R Markdown 6
Readr 6
Dplyr 7
Tidyr 7
Rpart 7
caTools 8
rpart.plot 8
GGPlot 9
TECHNIQUES 10
Shortcomings and Challenges 11
Outcomes 12
What Movie Rating Performs the Best at the Box Office? 12
What Movie Ratings (PG, G, R, etc.) have the highest average Quality Rating (1-10)? 13
What effect does a movie's Budget have on its Quality Rating? 15
What characteristics affect a movie's Quality Rating (1-10)? 16
Can a movie's Budget be used to identify its Rating (PG, R, G, PG-13, etc.)? 19
Confusion Matrix and Statistics 19
Confusion Matrix and Statistics 22
Does Runtime affect Popularity? 24
Conclusion 26
References 27

27

Executive Summary

The goal of this paper is to investigate a range of Business Intelligence (BI) techniques to identify what key factors are found in the most successful movies, and whether a movie can be a guaranteed success when it incorporates said key factors. Our research hopes to offer valuable information to stakeholders in the movie industry, enabling them to make more informed decisions and increase their chances of producing successful films.

Throughout this report, we provide a thorough explanation of our detailed dataset and the Business Intelligence (BI) processes that we utilized. We then describe the data mining tools that were applied to analyze the data and draw conclusions. Utilizing software tools like Microsoft Excel, R Studio, R Markdown, decision trees, linear regression, Readr, Dplyr, Tidyr, Rpart, caTools, rpart.plot and GGPlot, we were able to conduct a comprehensive analysis and create visual representations via bar charts and scatter plot graphs. The following sections of the report provide a detailed explanation of the graphs and their respective areas

Dataset Description

Our dataset consists of nearly 10,000 entries of successful movies, spanning across a one-hundred-year time span, however it should be noted that the vast majority of entries in our dataset are movies released within the twenty-first century. The criteria that we are measuring each movie with include movie title, budget, revenue, rating, runtime, genre, release date, and language among other things.

The data set contained a wide range of data, including:

- id: Unique identifier for each movie.
- imdb_id: Unique identifier for each movie, specific to the IMDB website.
- title: The title of a movie.
- budget: The given budget of a movie.
- revenue: The amount of money the film earned in revenues.
- certification: The given rating for a movie (G, PG, PG-13, R, NC-17).
- popularity: The given popularity based on IMDB's calculations.
- vote_count: The amount of votes a movie has on the IMDB website.
- vote_average: The average score a movie has based on votes on the IMDB website.
- runtime: The total length of a movie, measured in minutes.
- genres: The genre or genres that a movie is categorized in.
- release_date: The given day that a movie was released.
- release_year: The year that a movie was released.
- production_companies: The company or companies that produced the given movie.
- production_companies_country: The countries that the production companies operate from.
- language: The original language that a movie was released in.
- country: The country that a movie was released in.

The analysis offers a diverse range of data, providing valuable insights into various key factors for each of these movies. Through the utilization of data mining tools, significant data patterns

can be identified, leading to the development of effective business strategies, marketing analyses, knowledge bases, and other various insights.

Data provided in this paper consists of movies from the TMDB database (See Tables below)

Figure 1.1

id	imdb_id	title	budget	revenue	certification	popularity	vote_count	vote_average	runtime	genres	release_date	release_year
631842	tt15679400	Knock at the Cabin	20000000	52000000	R	3975.228	866	6.476	100	Horror, Mystery, Thriller	2/1/2023	2023
505642	tt9114286	Black Panther: Wakanda Forever	250000000	858535561	PG-13	3095.305	3902	7.347	162	Action, Adventure, Science Fiction	11/9/2022	2022
315162	tt3915174	Puss in Boots: The Last Wish	90000000	442000000	PG	2589.341	4339	8.413	103	Animation, Adventure, Comedy, Family	12/7/2022	2022
646389	tt5884796	Plane	25000000	51000000	R	2068.255	802	6.859	107	Action, Adventure, Thriller	1/12/2023	2023
677179	tt11145118	Creed III	0	100000000	PG-13	1920.868	191	7.139	116	Drama, Action	3/1/2023	2023
1058949	tt13614388	Little Dixie	0	0	R	1688.474	56	6.402	105	Thriller, Action, Drama, Crime	2/3/2023	2023
937278	tt7405458	A Man Called Otto	50000000	103842303	PG-13	1545.382	522	7.9	126	Comedy, Drama	12/28/2022	2022
772515	tt11718516	Huesera: The Bone Woman	0	0		1436.445	84	6.631	93	Drama, Horror, Mystery	2/10/2023	2023
1077280	tt26198528	Die Hart	0	0		1386.983	102	6.627	85	Action, Comedy, Thriller	2/22/2023	2023
842942	tt9853500	Bandit	0	0	R	1379.473	98	6.306	126	Crime, Drama, Thriller	9/23/2022	2022
536554	tt8760708	M3GAN	12000000	167643991	PG-13	1319.223	1915	7.496	102	Science Fiction, Horror, Comedy	12/28/2022	2022
76600	tt1630029	Avatar: The Way of Water	460000000	2281000000		1318.499	5631	7.739	192	Science Fiction, Adventure, Action	12/14/2022	2022
823999	tt15307230	Diabolik - Ginko all'attacco!	0	0		1308.043	71	5.796	111	Crime, Action	11/17/2022	2022
758009	tt9686790	Shotgun Wedding	0	4495365	R	1207.329	607	6.362	101	Action, Romance, Comedy	12/28/2022	2022
640146	tt10954600	Ant-Man and the Wasp: Quantum	200000000	419000000	PG-13	1032.078	913	6.407	125	Adventure, Science Fiction, Comedy	2/15/2023	2023
267805	tt1655444	There Are No Saints	0	0		952.788	66	5.553	105	Action	5/27/2022	2022
436270	tt6443346	Black Adam	200000000	392952111		841.29	4369	7.147	125	Action, Fantasy, Science Fiction	10/19/2022	2022
1035806	tt22394702	Detective Knight: Independence	0	0	R	818.828	81	6.012	92	Action, Crime, Thriller	1/20/2023	2023
785084	tt13833688	The Whale	3000000	32000000	R	799.218	830	8	117	Drama	12/9/2022	2022
842544	tt14873054	Transfusion	0	0		755.138	127	6.602	105	Crime, Thriller, Drama	1/5/2023	2023
906221	tt16280138	Magic Mike's Last Dance	45000000	52000000	R	746.568	107	7.098	112	Comedy, Drama, Romance	2/9/2023	2023
840789	tt12390388	Matadero	0	0		705.46	1	3	106	Drama	12/9/2022	2022
1003580	tt22769820	Legion of Super-Heroes	0	0	PG-13	691.55	60	6.692	84	Animation, Action, Science Fiction	2/7/2023	2023
676547	tt9271672	Prey for the Devil	0	37655699	PG-13	666.726	543	7.2	93	Horror, Thriller	10/23/2022	2022
653851	tt7693316	Devotion	90000000	20000000	PG-13	655.399	353	7.381	139	War, History, Drama	11/23/2022	2022
996727	tt15441472	The Price We Pay	0	0	R	630.425	84	6.1	85	Action, Thriller, Mystery, Crime, Horror	1/13/2023	2023

Figure 1.2

production_companies	production_companies_country	language	country
Blinding Edge Pictures, Universal Pictures, FilmNation Entertainment, Wishmore, Perfect World Pictures	US, US, US, US, CN	en	United States of America
Marvel Studios	US	en	United States of America
DreamWorks Animation, Universal Pictures	US, US	en	United States of America
MadRiver Pictures, Di Bonaventura Pictures, G-BASE, Olive Hill Media, Riverstone Pictures	US, US, US, US, GB	en	United States of America
Metro-Goldwyn-Mayer, Proximity Media, Balboa Productions, Outlier Society Productions, Chartoff-Winkl	US, US, US, US	en	United States of America
Roxwell Films, Three Point Capital	US, US	en	United States of America
Playtone, SF Studios, 2DUXÂ², Stage 6 Films, Artistic Films, STX Entertainment	US, SE, US, US, US	en	Sweden
Disruptiva Films, Machete Producciones, Maligno Gorehouse, Terminal	MX, MX, MX, MX	es	Mexico
HartBeat Productions	US	en	United States of America
Goldrush Entertainment, Yale Productions, Lucky 13 Productions	CA, US, US	en	Canada
Atomic Monster, Blumhouse Productions, Divide / Conquer, Universal Pictures	US, US, US, US	en	United States of America
20th Century Studios, Lightstorm Entertainment	US, US	en	United States of America
Mompracem, Astorina S.A.C., Emilia-Romagna Film Commission, RAI Cinema	IT, IT	it	Italy
Lionsgate, Mandeville Films, Nuyorican Productions, Maximum Effort	US, US, US, US	en	United States of America
Marvel Studios, Kevin Feige Productions	US, US	en	United States of America
Open Window Productions, BN Films, Itaca Films, Itaca Films	MX, US	en	United States of America
New Line Cinema, Flynn Picture Company, DC Films, Seven Bucks Productions	US, US, US, US	en	United States of America
Arcana Studio, 308 Enterprises, BondIt Media Capital, Buffalo 8	CA, CA, US, US	en	Canada
A24, Protozoa Pictures	US, US	en	United States of America
Deeper Water Films, Madman Entertainment, Altitude Media		en	Australia
Warner Bros. Pictures, Nick Wechsler Productions, Free Association	US, US, US	en	United States of America
El Viaje Films, 4 Ã 4 Productions, Prisma Cine, Magoya Films	ES	es	Argentina
Warner Bros. Animation, DC Entertainment, Warner Bros. Home Entertainment Group	US, US, US	en	United States of America
Lionsgate, Gold Circle Films, Confluence Productions	US, US	en	United States of America
Black Label Media, STX Entertainment, Columbia Pictures, Stage 6 Films	US, US, US, US	en	United States of America
828 Productions, Buffalo 8, 828 Media Capital, VMI Worldwide, ABS Payroll & Production Accounting Servi	US	en	United States of America

Tools and Techniques

TOOLS

Excel

Excel is an electronic spreadsheet software program developed by Microsoft. It allows users to organize, analyze, and manipulate data in a tabular format. Excel provides a grid of cells that can be filled with numbers, text, or formulas, and supports a wide range of mathematical and statistical functions for working with data. Excel is a powerful tool for exploring data. It provides a user-friendly interface for creating and working with spreadsheets, which are used to organize data into rows and columns.

R Studio

RStudio is an integrated development environment (IDE) for R, a programming language used for statistical computing and graphics. RStudio provides a user-friendly interface for working with R, making it easier to write, debug, and execute R code. It includes various features such as code highlighting, auto-completion, debugging tools, and package management, among others.

R Markdown

R Markdown is a file format for creating dynamic documents that combine text, code, and output. It is a simple and flexible way to produce reproducible reports, presentations, and websites from R code.

In an R Markdown document, plain text is written using Markdown syntax, which allows you to format text, add headers and sections, create lists and tables, and insert hyperlinks and images. You can also embed R code chunks within the document, which can be executed and produce output when the document is rendered. This allows you to seamlessly integrate data analysis, visualization, and reporting into a single document.

Readr

Readr is an R package that provides a fast and efficient way to read rectangular data files, such as CSV (Comma Separated Values), TSV, and fixed-width files, into R. The Readr package is particularly useful when working with large datasets, as it is optimized for speed and memory efficiency. It can automatically detect and parse the column types of data, which can save time and reduce errors when reading in data.

Dplyr

Dplyr is an R package that provides a set of tools for working with data frames in a fast and efficient way. It is a package designed for data science and analysis.

The dplyr package provides a set of functions that can be used to manipulate, summarize, and transform data in a data frame.

Here are some of the key functions provided by dplyr:

filter(): allows you to subset a data frame based on specified conditions.

select(): allows you to select specific columns from a data frame.

One of the key advantages of dplyr is that it is designed to be fast and memory efficient, making it particularly useful when working with large data sets.

Tidyr

Tidyr is an R package that provides a set of tools for cleaning and reshaping data. The tidyr package provides several functions for working with data that has multiple variables in each column or that is in a non-standard format.

key functions provided by tidyr:

gather(): allows you to convert data that is in a wide format (multiple variables in each column) into a long format (one variable per column).

spread(): allows you to convert data that is in a long format (one variable per column) into a wide format (multiple variables in each column).

One of the key advantages of tidyr is that it is designed to work seamlessly with the dplyr package, which provides functions for manipulating and summarizing data frames. Together, tidyr and dplyr provide a powerful set of tools for cleaning and reshaping data in R.

Rpart

rpart is an R package that provides a set of functions for building classification and regression trees using the Recursive Partitioning and Regression Trees (RPART) algorithm. It is a widely used package in machine learning and predictive modeling, risk assessment, and customer segmentation.

Classification and regression trees are a type of decision tree that recursively partitions a dataset into smaller subsets based on the values of the predictor variables. At each step of the partitioning process, the

algorithm selects the predictor variable that provides the best split, based on a specified criterion, such as the Gini index or information gain.

The rpart package provides several functions for building and visualizing classification and regression trees, including:

rpart(): builds an RPART model based on a dataset and a set of tuning parameters.

printcp(): displays the complexity parameter table for an RPART model.

prp(): plots the decision tree for an RPART model in a compact and informative way.

One of the key advantages of rpart is that it is computationally efficient and can handle large datasets with many predictor variables. It also provides a flexible framework for building decision trees, with several options for controlling the size and complexity of the trees.

caTools

The caTools package is an R programming language library that provides various utility functions for data manipulation. The package includes a wide range of functions, such as reading and writing data, generating summary statistics, and manipulating data frames.

sample.split(): This function is used to split data into training and testing sets. It randomly splits the data into two sets based on a specified proportion.

colSplit(): This function is used to split a data frame into columns. It takes a data frame and a delimiter as inputs, and returns a list of vectors

rpart.plot

rpart.plot is an R package that provides a set of functions for visualizing decision trees created using the rpart package. It is a popular tool for creating clear and informative plots of decision trees. The rpart.plot package provides several functions for visualizing decision trees, including:

rpart.plot(): creates a detailed and customizable plot of a decision tree created using the rpart package. One of the key advantages of rpart.plot is that it allows you to create customized plots of decision trees that are easy to understand and interpret. The plots can be customized in a variety of ways, including changing the size, color, and shape of the nodes and branches, and adding text labels and annotations.

GGPlot

ggplot is a data visualization package in the R programming language that is widely used for creating high-quality graphics. The package is based on Grammar of Graphics, which is a framework for creating graphics that allows you to create complex plots with simple code. Wide variety of visualizations, including scatterplots, bar plots, line graphs, histograms, and more. The package allows you to customize every aspect of your plot, from the axes and labels to the colors and fonts.

TECHNIQUES

An API (Application Programming Interfaces) endpoint for The Movie Database (TMDb) was used to provide access to TMDb's movie data, allowing for the creation of applications that can retrieve and display movie information about a movie with the ID 76341.

The API_key parameter is used to authenticate the API request and ensure that the request is coming from a valid source. The API_key is a unique identifier issued to developers when they register for access to the TMDb API. It is used to authenticate each API request, ensuring that only authorized users can access the data. To use this API endpoint, we replaced <<API_key>> with our own unique API key provided by TMDb and we were able retrieve movie information and dataset from TMDb.

We use Excel to study and understand the dataset to determine the number of records and rows, the dataset consists of 28 records and 10,000 rows.

The R markdown was used for the analysis of this project. The specific libraries, packages, and functions used for data manipulation, visualization, modeling, and analysis, are as follows Readr, dplyr, tidyr, rpart, rpart.plot, ggplot contributed to the success of the project. These packages provided us with a set of functions to efficiently clean, reshape, and transform data.

For visualization, ggplot2 and rpart.plot package were used to create highly customizable and aesthetically pleasing plots. We also used the Plotly package to create interactive plots that could be shared and explored online. During modeling and prediction, we used several packages, including Caret, and CaTools. These packages provided us with various machine learning algorithms, such as decision trees, random forests, and other forecasting methods.

For data analysis, packages such as, dplyr, tidyr and caTools were used, these packages perform descriptive and inferential analyses, including hypothesis testing, regression analysis, text analysis, and financial analysis.

Overall, the combination of these tools and techniques allowed us to work efficiently with R to derive meaningful insights from data.

Shortcomings and Challenges

While TMDB (The Movie Database) API is a great resource for movie enthusiasts and researchers, it does have some shortcomings and limitations. It has limited information on certain titles. While TMDB has a vast collection of movies and TV shows, some titles may not have complete or update-to-date or accurate information available, such as missing budget, revenue, quality rating, rating, release date, and/or release year.

There is a risk of inaccurate or incomplete data because TMDB relies on 100% user information contribution. They encourage the users to add missing items themselves and edit any incorrect data that users find. This can be particularly problematic when it comes to the accuracy of ratings, as it can be skewed by individuals with interests or biases. TMDB has a discussion forum for users to engage with each other. This can help users to get a sense of what titles are popular and get more information on content or skew other individuals' thoughts. TMDB has limited access to industry data. Unlike some other movie databases, TMDB does not provide comprehensive information on box office revenue which can be important for industry or college student analysts.

TMDB API imposes a limit on the number of requests that can be made in each time frame. This limited the amount of data the team retrieved from TMDB API. The team wanted to get access to a large amount of data to provide the most accurate presentation of the dataset.

TMDB API lacks search and filtering options, especially for more niche or specialized genres. This can make it difficult to find specific types of content. The layout of the content can be difficult to work with even though TMDB API offers a lot of functionality and data access.

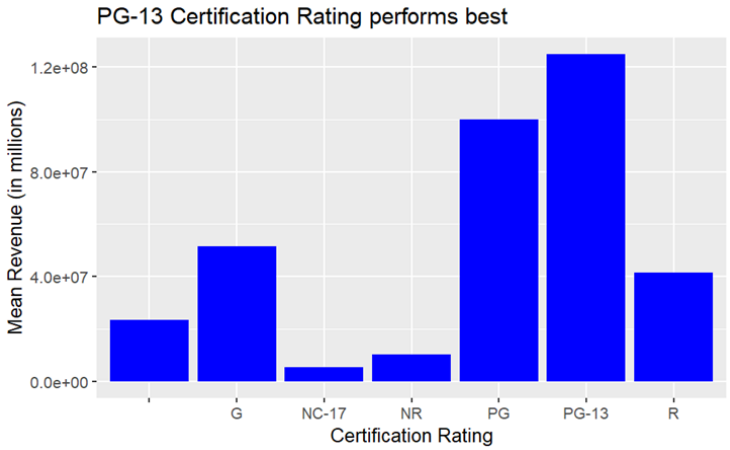
TMDB API has limited data access. While TMDB API provides access to a lot of data, some important information was not available through the API. This limitation of search options caused difficulty retrieving information from TMDB API. One of the limitations was that it did not support searching by multiple criteria or filtering results by specific attributes.

There is documentation and support resources for TMDB API, but they are not complete or not fully up to date and they are hard to use. The documentation provides basic information about the TMDB API such as creating the end points for a simple output. It provides the database's structure, but there are few nested attributes that were hard to access, such as rating. This made it challenging for the team to get up and running quickly with the API.

Outcomes

What Movie Rating Performs the Best at the Box Office?

To answer the question what movie rating (PG, R, G, PG-13, etc.) performs the best at the Box Office (i.e., highest revenue), we plotted a graph to show this.

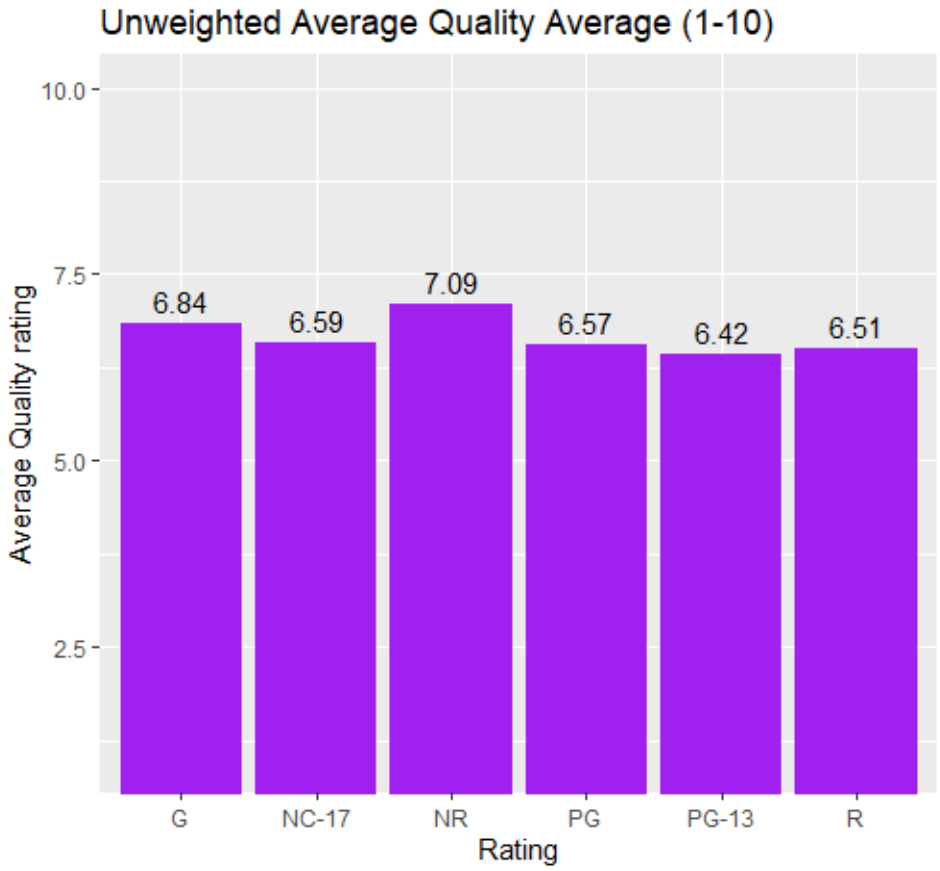


From the above chart it shows that the rating that performs best at the box office is PG-13. The result also provides an insight showing how budget is a key influencing factor in movie production quality. Movies with higher budget such as PG-13 have a higher revenue, with PG also having a good rating because it comes second.

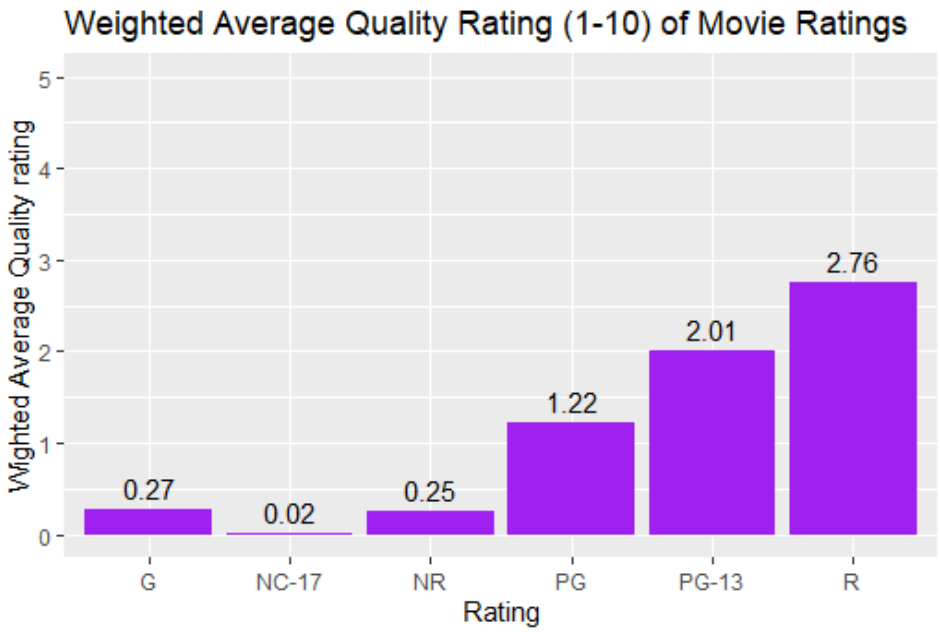
NC-17 and NR are the least rated, this shows that both have lower budgets and revenue which indicates poor movie quality and performance according to the graph.

What Movie Ratings (PG, G, R, etc.) have the highest average Quality Rating (1-10)?

After reviewing the performance of the various film Ratings at the box office, we thought an important insight into the data would be to determine which films are rated the highest on average? We determined best that we use both weighted and unweighted averages of the data due to the varying representation size of each film rating in our data set. For example, NC-17, a film rating that does not permit any viewer under the age of 17 from watching, is quite rare and appears less than 10 total times in the dataset.



Here, we see NR (Not-Rated) films taking the highest average while NC-17 has the lowest. You should notice right away that there is not much variance among the films in their current state. To better show how the films are represented against one another, a weighted average was taken and shown below.



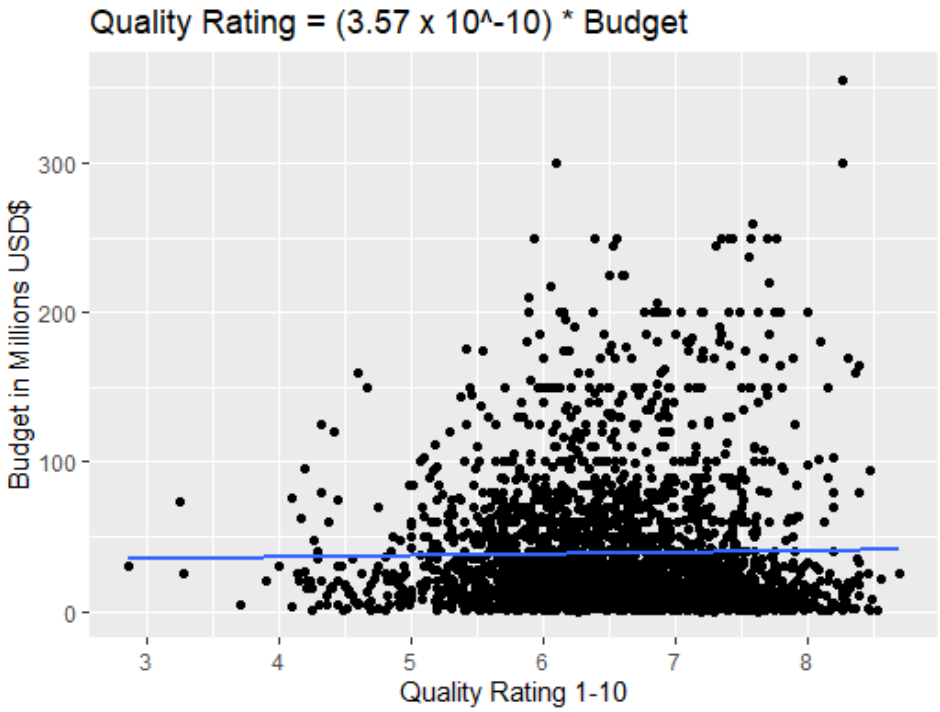
Here, the variances differ significantly more than their unweighted counterparts. In this case, the weighted average is significantly lower in each Rating than before. This is because by solving the weighted average, we multiply each variable by its rating's respective proportion of the total dataset, then divide it by the weighted sum of all the films. This just signifies that a lot of the films are scored around the numbers most of the time, however, scores for one rating that are farther away from the averages could be rarer than for the other ratings.

For example, take NR rated films. NR rated films have the highest unweighted rating but the 2nd lowest weighted rating. This would mean that there are far fewer high-scoring films than low-scoring films for NR films. Thus, you are bound to see more NR films that are considered not good critically than you would otherwise.

What effect does a movie's Budget have on its Quality Rating?

Studios spend a wide-ranging amount of money on films in the hope to make the best possible motion picture. This is not always the case. For our project, we wanted to analyze the effect of movie budgets on the quality rating of the same film. To do so, we ran a simple linear regression on the two variables with the quality rating being the dependent variable and budget being the dependent variable. Below, you will see the resulting graph from the regression along with the regression equation in the title along with a summary report of the regression.

```
# Null Hypothesis: A Movie's Budget does not have an effect of its Quality Rating.
# Alternative Hypothesis: Movie characteristics have an effect of its Quality Rating.
# Use 95% confidence:
# Alpha = 0.05
```



As seen through the high p-value of 0.2895, it is significantly greater than our alpha value of 0.05, thus we reject the Null hypothesis that Budget affects a film's quality rating. You can also determine that budget is not statistically significant in determining the quality rating of a film. While we can determine that budget is not an indication of a film having a good or bad quality rating, this does beg the question of what affects quality rating?

What characteristics affect a movie's Quality Rating (1-10)?

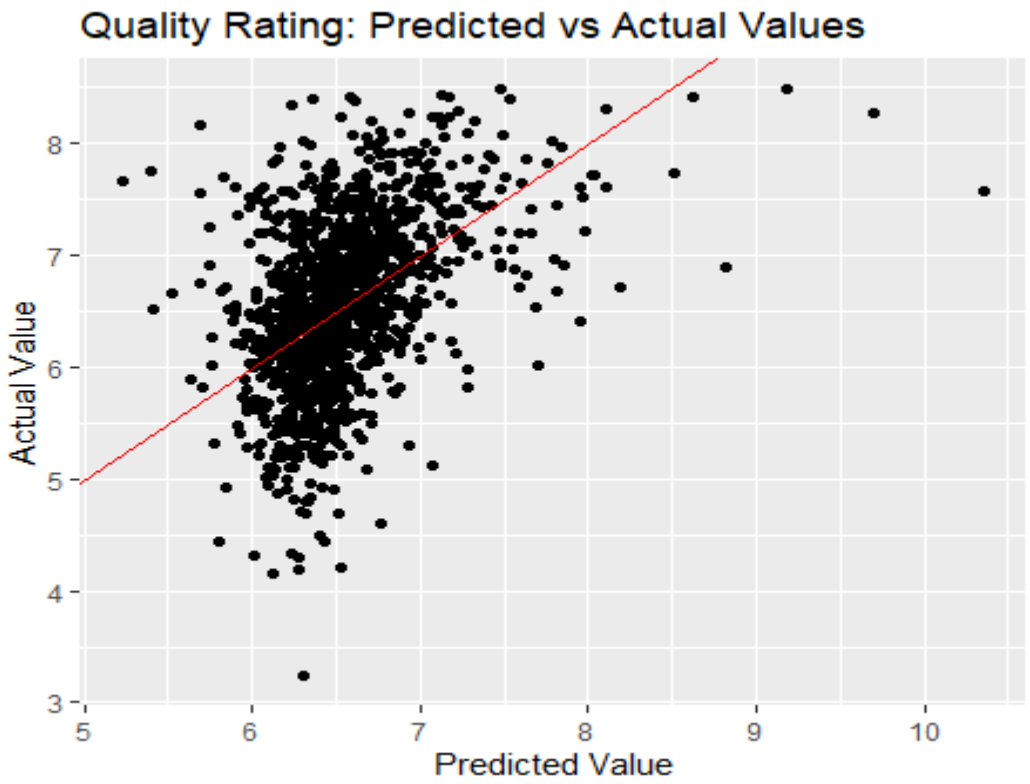
To determine the effect of a film's other characteristics on its quality rating, a multiple linear regression model was created using its budget, revenue, runtime, and title length. Below, you will see a summary of the initial results. We will also use a confidence level of 95%, giving us an alpha value of 0.05.

```
Call:
lm(formula = quality_rating ~ budget + revenue + title_length +
    runtime, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2665 -0.4208 -0.0028  0.4443  2.4274

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.898e+00  6.297e-02  77.777   <2e-16 ***
budget      -6.124e-09  3.519e-10 -17.406   <2e-16 ***
revenue      1.399e-09  8.107e-11  17.260   <2e-16 ***
title_length 8.409e-04  1.244e-03   0.676    0.499
runtime      1.545e-02  5.512e-04  28.026   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6866 on 4101 degrees of freedom
Multiple R-squared:  0.2273,    Adjusted R-squared:  0.2266
F-statistic: 301.7 on 4 and 4101 DF,  p-value: < 2.2e-16
```

We can determine that all variables are very significant except for title length, which is not significant at all. We then remove title length as a variable and rerun the regression to produce the following summary of results.

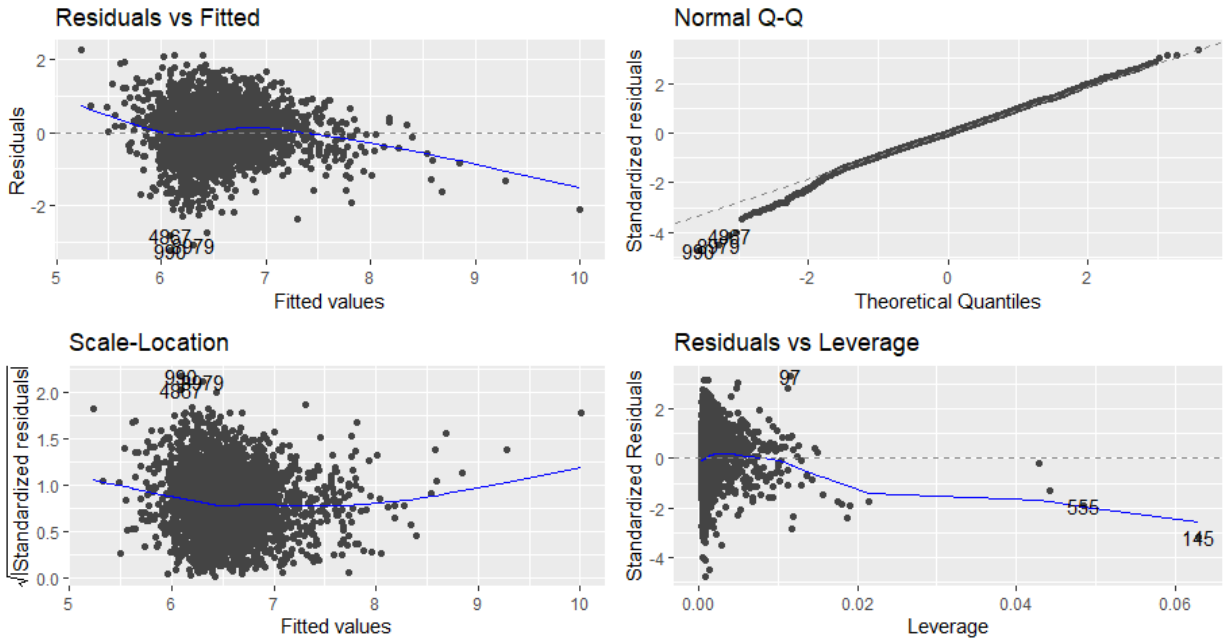
```
Call:
lm(formula = quality_rating ~ budget + revenue + runtime, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2338 -0.4150 -0.0061  0.4455  2.2602

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.813e+00  7.175e-02   67.08   <2e-16 ***
budget      -6.550e-09  4.261e-10  -15.37   <2e-16 ***
revenue      1.523e-09  1.021e-10   14.92   <2e-16 ***
runtime      1.633e-02  6.639e-04   24.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6834 on 2856 degrees of freedom
Multiple R-squared:  0.2429,    Adjusted R-squared:  0.2421
F-statistic: 305.4 on 3 and 2856 DF,  p-value: < 2.2e-16
```

The model's p-value is much smaller than our alpha value of 0.05, so we begin to move on by creating a prediction of values and plotting them with the actual values. The red line shows the slope of the line that the values should follow if they are all completely accurate.



We can determine just from that graph that the model does a very poor job at correctly predicting a film's quality rating with a very low number of values being correctly scored. In fact, when calculating the prediction's error rate, we found evidence to support this. The root mean squared error (RMSE) and residual standard error (RSE) were considered here. RMSE is the average error performed by a model for a given observation. Our RMSE for the multi-linear regression model above was 69.54%. RSE is a variant of RMSE adjusted for the predictors (independent variables) in the model. The RMSE came out to be a whopping 69.54% while the RSE was only 10.48%. Typically, in multi-linear regression, we want these values to be close together and be low. The large difference between the two indicates the high amount of error and failure of the model.

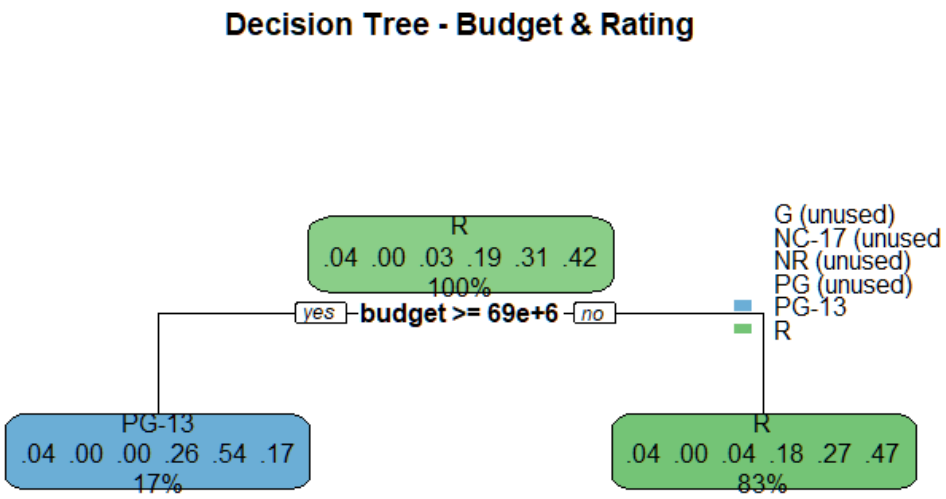
Using the residual plots below, we can further determine the validity of the regression model. You can see in the Residuals vs. Fitted Graph that the values fan out, thus indicating that a linear model is not the best choice for modelling this data. This reinforces the high error rates for RMSE and RSE earlier.



While these results do show inaccurate predictions, they do provide some insights as there is a lot of variability in film making and success. For example, there have been several films that have done very well critically (I.e., high quality rating), but very poorly at the box office (I.e., low revenue) and vice versa. It is extremely difficult to predict the critical score of a movie as there are several factors that play into the score that are not always numerical or statistically driven, such as personal preference, bias, source of content, and others. Moving forward, it would be a good indication to use multi-linear regression in a similar way to the decision trees that are used in the next section: to indicate if a film's quality rating is at all affected by its rating (PG, G, R, NR, PG-13, etc.). However, in the end, it is apparent that regression cannot consistently and successfully predict a film's quality rating based on the results above.

Can a movie's Budget be used to identify its Rating (PG, R, G, PG-13, etc.)?

R was used to attempt to analyze the relationship between a movie's budget and its quality rating by building a decision tree model on a dataset of movies with their respective budgets and ratings. The data is then split into training and testing sets using the sample.split() function from the caret package. The split ratio is set to 0.7, which means that 70% of the data was used for training and 30% for testing. A decision tree is then built on the training data using the rpart() function from the Rpart package. The formula rating~. specifies that the rating column is the dependent variable, and all other columns in the data frame are independent variables. The resulting decision tree is displayed using the rpart.plot() function.



The decision tree model is then used to predict the ratings of the movies in the testing set using the predict() function. The type of prediction is set to "class," meaning the predicted ratings will be categorical rather than continuous.

The confusionMatrix() function from the Caret package is then used to create a confusion matrix of the predicted and actual ratings. The confusion matrix shows the number of true positives, true negatives, false positives, and false negatives for each rating category.

Confusion Matrix and Statistics

Reference

Prediction	G	NC-17	NR	PG	PG-13	R
G	0	0	0	0	0	0
NC-17	0	0	0	0	0	0
NR	0	0	0	0	0	0
PG	0	0	0	0	0	0
PG-13	14	0	2	97	175	60
R	65	6	76	272	469	825

Overall Statistics

Accuracy : 0.4852

95% CI : (0.4634, 0.507)

No Information Rate : 0.4294

P-Value [Acc > NIR] : 0.0000001943

Kappa : 0.128

In the confusion matrix, the model predicted the class 'PG-13' for 348 instances, of which 175 were 'PG-13' (true positives). The model also predicted 'R' for 1713 instances, of which 825 were 'R' (true positives). The model did not predict any other class correctly.

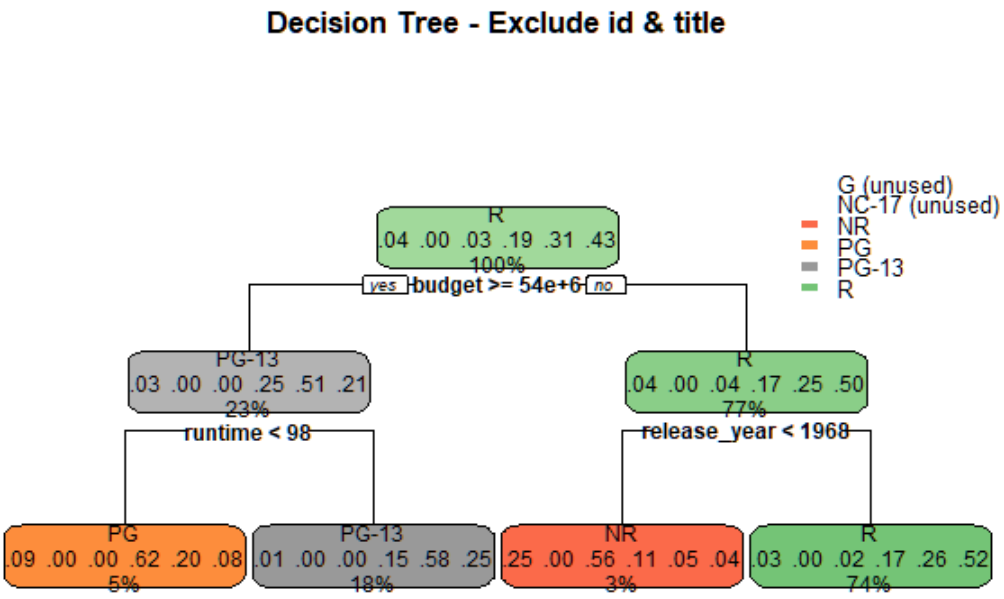
The no-information rate (NIR) is 42.94%, which is the accuracy we would get by always predicting the most frequent class, which in this case is 'R'. The p-value for the accuracy being greater than the NIR is 0.0000001943, which is less than 0.05 and indicates that the model is significantly better than the most frequent class. This suggests that the model has learned some useful patterns in the data and is making predictions better than what would be expected by chance.

The overall accuracy of the model is reported as 0.4852. Which means that the model correctly classified 48.52% of the data point. The range given, between 0.4634 and 0.507, indicates that the true accuracy

of the model is estimated to be within this range with 95% confidence. Kappa coefficient is 0.128, which the value is less than 0.2. This Indicates poor agreement between the predicted and actual ratings.

The decision tree model does not seem to be a good fit for predicting the ratings of movies based on their budgets. The confusion matrix and statistics suggest that the model is not performing well, as it is only correctly classifying less than 50% of instances and has poor agreement with the actual classes. Moreover, the model is only predicting two out of the six classes 'PG-13' and 'R' and is not performing well for any other class.

The same process was used when more variables were introduced to the model. All variables are used except id and title. The resulting decision tree is as follows.



Confusion Matrix and Statistics

Reference

Prediction	G	NC-17	NR	PG	PG-13	R
G	0	0	0	0	0	0
NC-17	0	0	0	0	0	0
NR	16	0	33	11	3	2
PG	8	0	0	39	19	2
PG-13	4	0	1	48	151	78
R	34	6	34	181	323	552

Overall Statistics

Accuracy : 0.5016

95% CI : (0.4764, 0.5269)

No Information Rate : 0.4104

P-Value [Acc > NIR] : 0.0000000000002839

Kappa : 0.2108

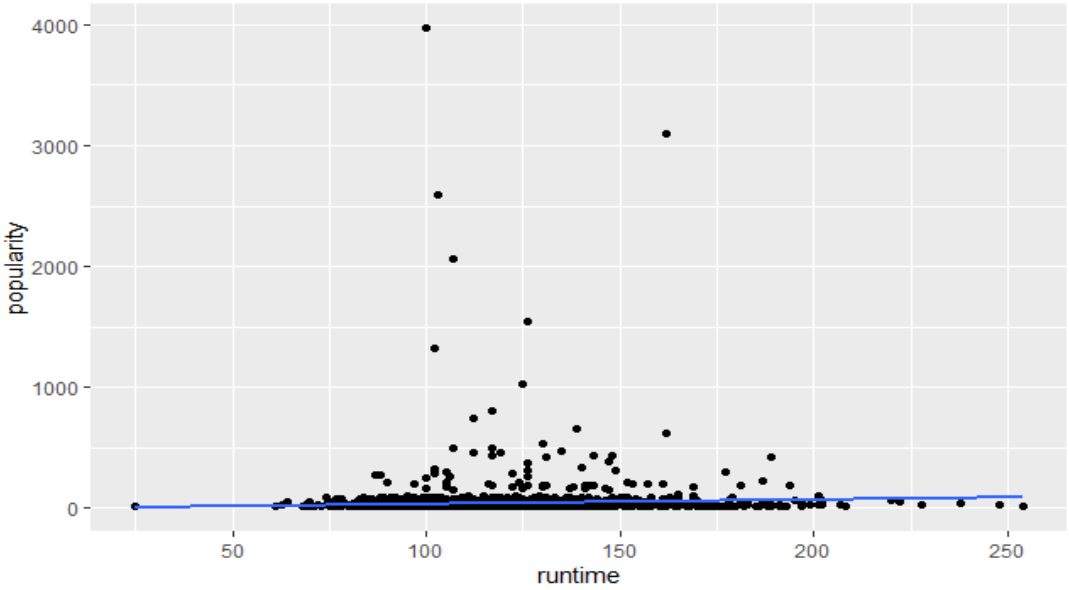
The accuracy of the model is 0.5016, which means the model predicts 50.16% of the movies correctly. The Kappa score of 0.2108 indicates that the model's performance is not much better than chance. The 95% confidence interval for accuracy is (0.4764, 0.5269), which means that if we repeat the experiment many times, we can expect the accuracy to lie in this range 95% of the time.

Looking at the statistics by class, we can see that there were no instances predicted as 'G' or 'NC-17'. For 'NR' and 'PG' only around 30 were predicted. 'PG-13' was correctly classified 151 times. We can see that the highest value in the diagonal is 552, which corresponds to class 'R'. Therefore, the most frequent class in terms of the number of instances that were correctly classified is class 'R'.

The no-information rate (NIR) is 0.4104 and the most frequent class is 'R'. The p-value for the accuracy being greater than the NIR is 0.0000000000002839, which is less than 0.05 and indicates that the model is significantly better than the most frequent class. The kappa statistic is 0.2108 and a kappa value between 0.2 and 0.4 indicates fair agreement.

Overall, the decision tree model does not perform well, as it only correctly classifies half of the observations. The confusion matrix and statistics suggest that the model is performing better than the most frequent class, but its performance is still not satisfactory. The accuracy of 0.5016 is better than the no-information rate of 0.4104, but it still indicates that the model is correctly classifying only around 50% of the instances. The kappa value of 0.2108 indicates fair agreement between the model's predictions and the actual classes, but it is still not high enough to suggest a strong performance.

Does Runtime affect Popularity?



This output is the result of a linear regression model fitted to a dataset with two variables: "popularity" and "runtime". The goal of linear regression is to determine if there is a relationship between the two variables.

The first part of the output shows the residuals, which are the differences between the predicted values of popularity and the actual values of popularity for each data point in the dataset. The residuals help to assess the accuracy of the linear regression model.

The second part of the output shows the coefficients of the linear regression model. The intercept represents the expected value of popularity when runtime is zero, which may or may not be meaningful depending on the context of the data. The coefficient for runtime represents the change in popularity for each unit increase in runtime. Both the intercept and coefficient have corresponding standard errors, t-values, and p-values, which are used to assess the significance of the estimates.

The last part of the output shows various statistical measures of the model fit. The residual standard error represents the average difference between the predicted and actual values of popularity, and the multiple R-squared value represents the proportion of variability in the response variable (popularity) that can be explained by the predictor variable (runtime). The adjusted R-squared value is a modified version of the R-squared that adjusts for the number of predictor variables in the model. The F-statistic and corresponding p-value test the overall significance of the linear regression model, which is based on the null hypothesis that all coefficients are zero.

```
Call:
lm(formula = popularity ~ runtime, data = df)

Residuals:
   Min     1Q Median     3Q    Max
 -72.0  -17.9  -11.2   -0.5 3943.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.33389    9.25070  -0.468    0.639
runtime      0.36241    0.08296   4.368 1.28e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 107.2 on 4104 degrees of freedom
Multiple R-squared:  0.004628,  Adjusted R-squared:  0.004386
F-statistic: 19.08 on 1 and 4104 DF,  p-value: 1.283e-05
```

The model summary shows that the p-value of 2.656e-05 is exceedingly small, which indicates that the model is statistically significant and that there is evidence of a relationship between popularity and runtime.

Conclusion

While several aspects of our analysis ended up not being successful, there are still several key insights that were made into analyzing film data moving forward. The quick general analysis of the varying movie ratings and their performance both critically and at the box office gives insight as to the most common and most sought-after films on the market. Through the linear regression that was done, we were also able to determine that the dataset as constructed is not best modelled through a linear means. This work highlights the variability in film making and how even if you stick to a given formula, it will not perform as you may have hoped. It would be possibly more beneficial to apply non-linear models as well as make improvements to the database. These improvements could include items such as cast, directors, types of equipment used, shooting locations,

Another thing to consider with this analysis is the issues with TMDb. Several films while appearing in the database, had little to no data associated with them, including popular films from several years in the past, such as Star Wars and Jaws, even though there is plenty of accurate and up-to-date figures on the performance and characteristics of these films online. Given this, several films had to be discarded from the calculations because of this. It is possible that the inclusion of these films could skew findings slightly to some of the high-performing films high volume. This issue was more than likely created due to the API for the database being open-source and relying on user input to update figures characteristics of the films.

References

French, Ted. "Learn What Ms Excel Does Plus." *Lifewire*, Lifewire, 11 June 2020, https://www.lifewire.com/what-is-microsoft-excel-3573533.

"The Most Trusted IDE for Open Source Data Science." *Posit*, Posit Software, PBC, 18 Apr. 2023, https://posit.co/products/open-source/rstudio/.

"The Movie Database API." *Api Docs*, Themoviedb.org, https://developers.themoviedb.org/3/getting-started/introduction.

"Package Catools." *CRAN*, Comprehensive R Archive Network (CRAN), https://cran.r-project.org/web/packages/caTools/index.html.

"R Markdown." *R Markdown*, RStudio, PBC, https://rmarkdown.rstudio.com/.

"The R Project for Statistical Computing." *R*, The R Foundation, https://www.r-project.org/.

Therneau, Terry, et al. "Package Rpart - Cran.r-Project.org." *Rpart: Recursive Partitioning and Regression Trees*, Comprehensive R Archive Network (CRAN), 21 Oct. 2022, https://cran.r-project.org/web/packages/rpart/rpart.pdf.

Wickham, Hadley, et al. "A Grammar of Data Manipulation." *A Grammar of Data Manipulation* •, Tidyverse.org, https://dplyr.tidyverse.org/.

Wickham, Hadley, et al. "Create Elegant Data Visualisations Using the Grammar of Graphics." *Create Elegant Data Visualisations Using the Grammar of Graphics* •, Tidyverse.org, https://ggplot2.tidyverse.org/.

Wickham, Hadley, et al. "Tidy Messy Data." *Tidy Messy Data* •, Tidyverse.org, https://tidyr.tidyverse.org/.