Hyperiondev

TASK

Exploratory Data Analysis on the Movies Data Set

Visit our website

Introduction

The purpose of this report is to conduct an exploratory data analysis (EDA) of the 'movies' dataset, delving into its repository of information on 4803 movies. Through comprehensive data exploration and visualization techniques, we aim to glean valuable insights into the intricate dynamics of the film industry, uncovering trends, preferences, and factors influencing the success and reception of movies.

By leveraging the vast array of attributes provided in the dataset (budget, genres, popularity, release date, revenue, runtime, spoken languages, title, vote average and vote count), we seek to unravel the underlying narratives within the world of cinema. From identifying the most profitable genres to understanding the relationship between movie runtime and revenue, this report endeavours to offer a holistic perspective on the multifaceted landscape of movies.

Through our data-driven approach, we endeavour to shed light on the evolving preferences of audiences, the strategies employed by filmmakers and production companies, and the overarching trends shaping the cinematic landscape.

DATA CLEANING

To clean the ‘movies’ dataset, two methods were employed to enhance its relevance and accuracy for analysis:

1. Removing Unnecessary Columns:

Initially, unnecessary or redundant columns that were not utilized in the data analysis were identified and removed. This process ensured that only relevant information remained, making it easier to derive meaningful insights.

2. Duplicate Row Removal:

Duplicate rows were identified and removed from the dataset. This step was crucial in enhancing model efficacy and preventing bias, as duplicate data could skew the results of the analysis.

3. Date Formatting and Year Extraction:

The data type of the 'release date' column was converted to DateTime format to enable effective temporal analysis. This conversion facilitates the exploration of trends and patterns based on the release dates of movies. Additionally, a separate column was created to extract the year information from the 'release date' column

. This allows for better examination of temporal trends within the dataset, enabling movie insights over time.

By implementing these cleaning methods, the dataset was refined to ensure relevance, accuracy, and optimal performance for subsequent analysis and modeling tasks.

MISSING DATA

1. Removal of missing values

To address missing data, movies with a budget or revenue amount of 0 were deemed as missing values . They could not be accurately estimated so they were excluded from the dataset because as there is not a reliable method to substitute this missing information. This missing data falls under the category of missing completely at random (MCAR), indicating that it's not linked to any observed or unobserved factors in the dataset. The cause of this missing data is entirely unknown and removing it would not introduce any significant bias or distortion to the dataset.

DATA STORIES AND VISUALISATIONS

Valuable insights were gained from the data by creating and analyzing visual representations. Below are the various data stories along with their visualizations:

Top 5 most expensive and cheapest movies

Figure 1 is a bar plot comparing the budgets for the five most expensive movies and the five cheapest movies in the dataset. This comparison offers a straightforward visual depiction of the budget diversity within the dataset, showcasing both the highest and lowest budget allocations. It provides insight into the significant gap between high-budget blockbuster productions and lower-budget films, aiding in the comprehension of budget distribution across different movie types.

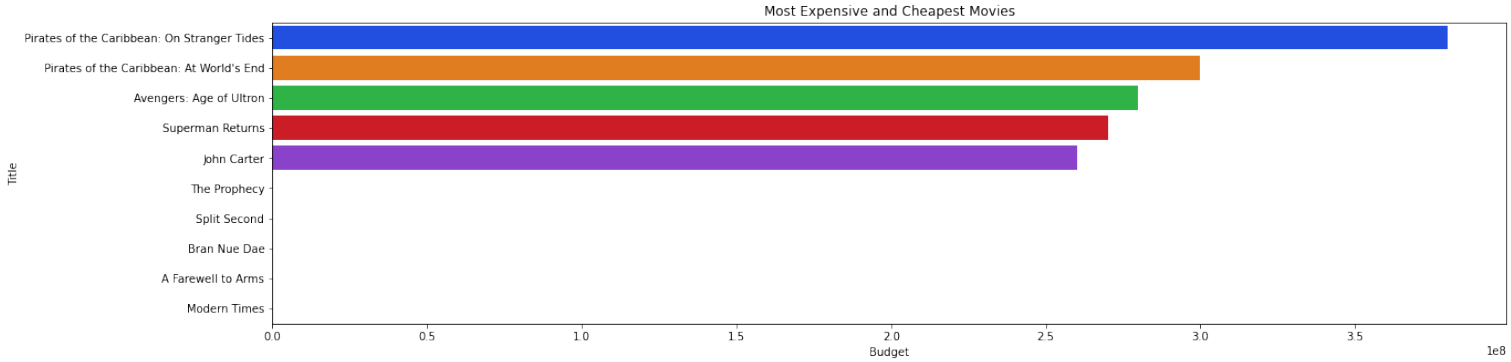

Figure 1

Top 5 most and least profitable movies

Figure 2 is a bar plot comparing the top five most profitable movies with the five least profitable ones. By calculating the difference between the revenue generated and the budget spent, we can determine the profit earned. This comparison allows us to distinguish between successful and unsuccessful approaches in movie production. Notably, it is evident that the five movies with the lowest profits actually resulted in financial losses.

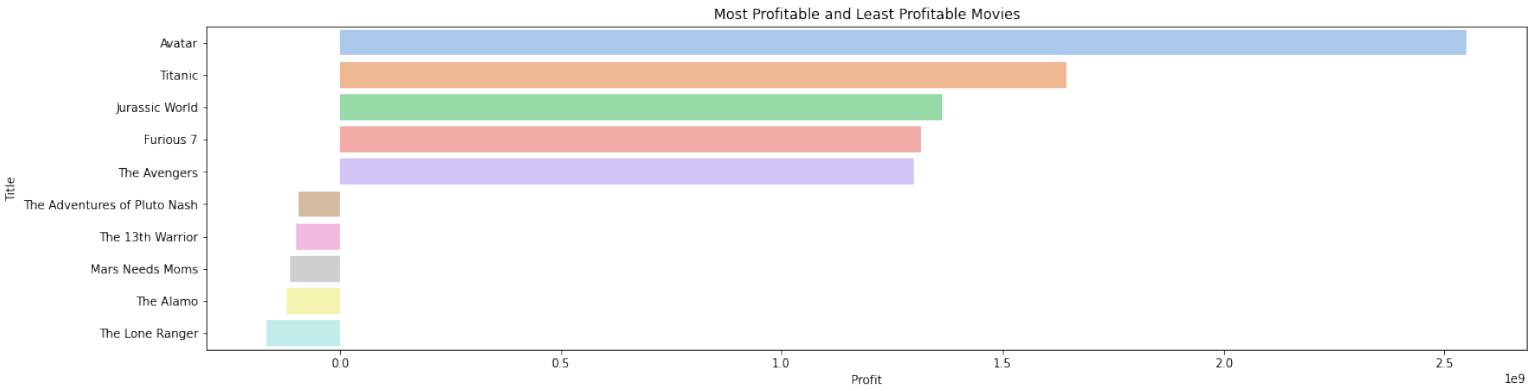

Figure 2

The most popular movies

Figure 3 is a bar plot that ranks the top 15 most popular movies based on their popularity rating, with the most popular movie positioned at the top and popularity decreasing as you move towards the bottom. This visualization offers valuable insights into consumer preferences, trends, and behaviours in the realm of movies.

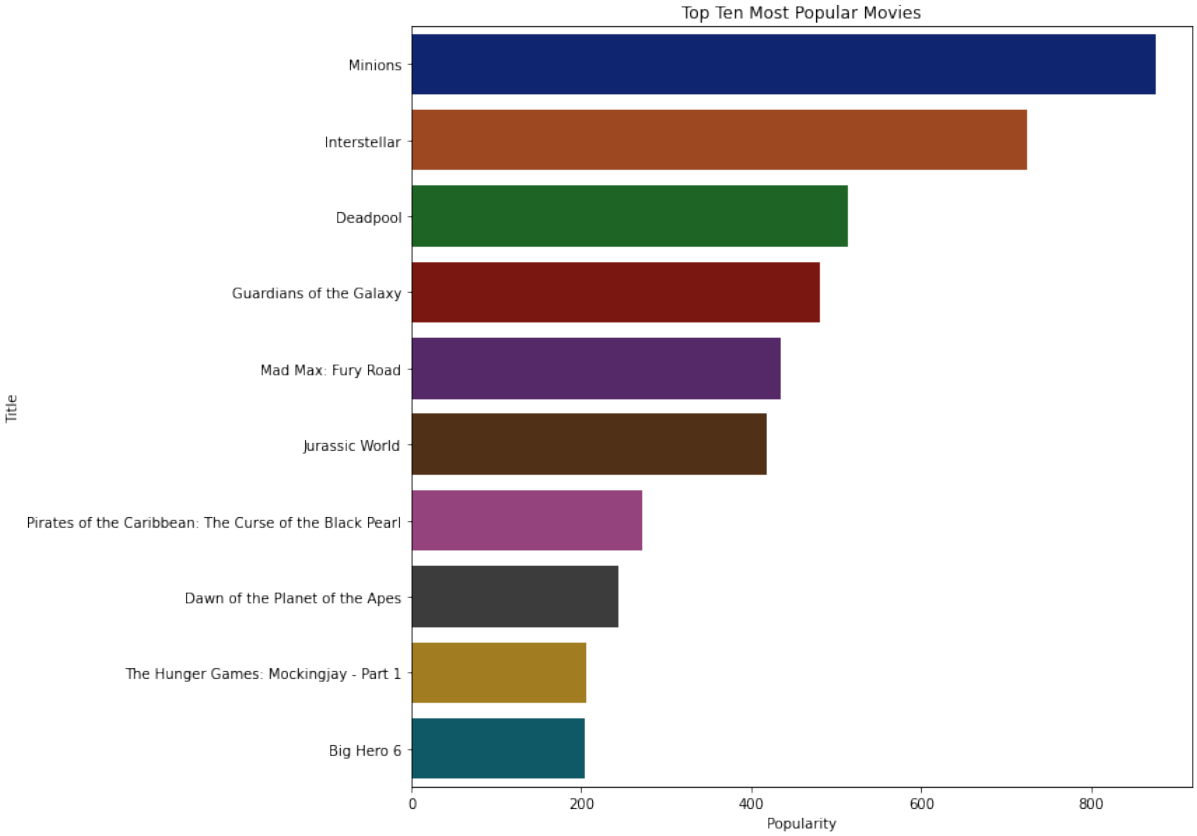

Figure 3

Movies rated above 7

Out of all the movies in the dataset, 637, which makes up 13.3%, are rated above 7. *Figure 4* shows the top 15 movies with a rating above 7. It is probable that when consumers aim to select a worthwhile movie, they may prefer choosing from this selection with ratings above 7. This helps ensure a satisfying experience and alleviates the pressure of picking something good

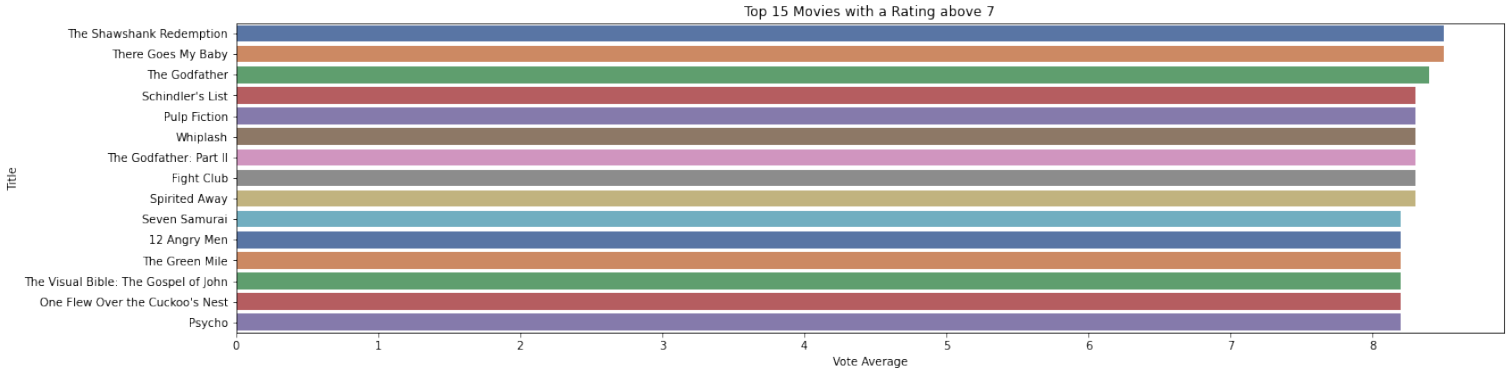

Figure 4

Most successful genres

Figure 5 is a bar chart illustrating the frequency of movies in each genre, arranged with the highest frequency on the left and the lowest on the right. It's notable that Drama is the most frequent genre. According to (Screencraft, 2024), this preference for Drama among movie producers could be attributed to several reasons:

- Dramas often explore relatable themes such as love, loss, and identity, which resonate with audiences globally.

- They tend to receive accolades and awards like the Oscars, enhancing their reputation and encouraging more dramas to be produced.

- Drama can be blended with other genres, offering diverse storylines while still captivating viewers with compelling characters and plots.

- Dramatic narratives prioritize character depth and emotional resonance, providing actors with opportunities to showcase their talents.

- They delve into real-life experiences and emotions, making them highly immersive and engaging.

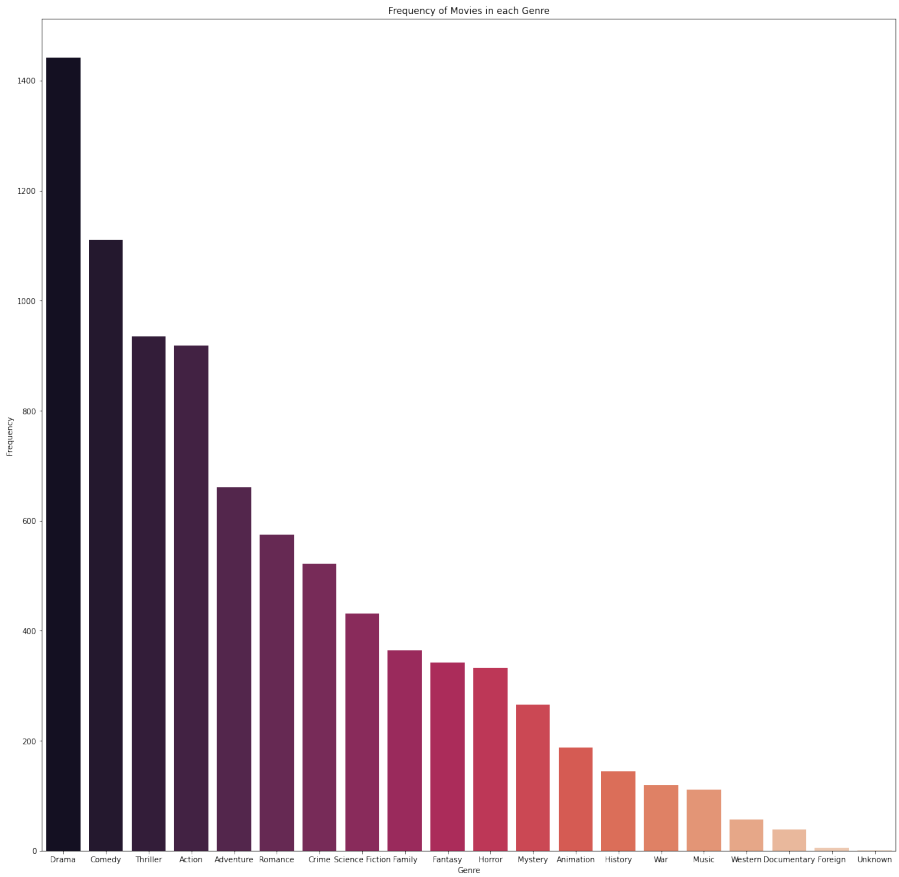

Figure 5

Genre that generates the most revenue

Figure 6 is a bar chart displaying the revenue generated by each movie genre, arranged with the highest revenue on the left and the lowest on the right. Notably, the Animation genre generates the most revenue. According to (StoneSoup, 2014) this trend can be attributed to several factors:

- Animation films appeal to a broad audience, ranging from children to adults, which increases ticket sales and merchandise revenue.

- They are often watched repeatedly by children, contributing to higher box office earnings.

- Animation movies tend to perform well globally as they are less impacted by language and cultural barriers.

- They generate additional revenue through merchandise such as toys, games, and clothing, further boosting their earnings.

- Successful animation films frequently lead to sequels or spin-offs, extending their revenue potential and sustaining profitability over time.

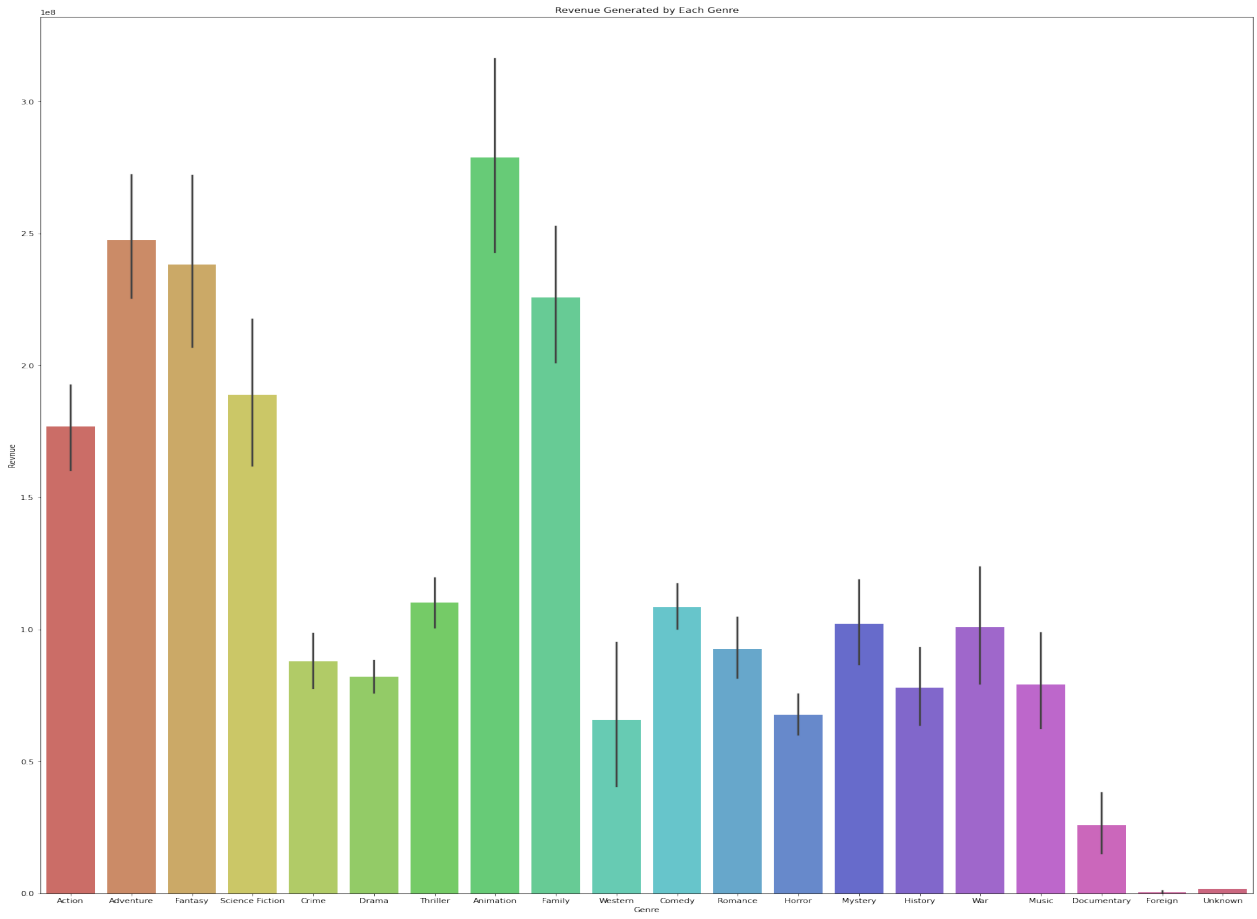

Figure 6

Year in which the most revenue was generated for 10 year span (2000 -2016)

Figure 7 presents a box and whisker plot illustrating the revenue generated by movies per year over a span of 10 years, from 2006 to 2016. Notably, the year 2014 stands out with the highest revenue. Additionally, the plot indicates an overall upward trend in movie revenue over time. This increase in revenue over time could be attributed to the overall, the upward trend in movie revenue suggests a positive trajectory for the film industry, fuelled by factors such as technological advancements, audience expansion, and evolving consumer behaviours.

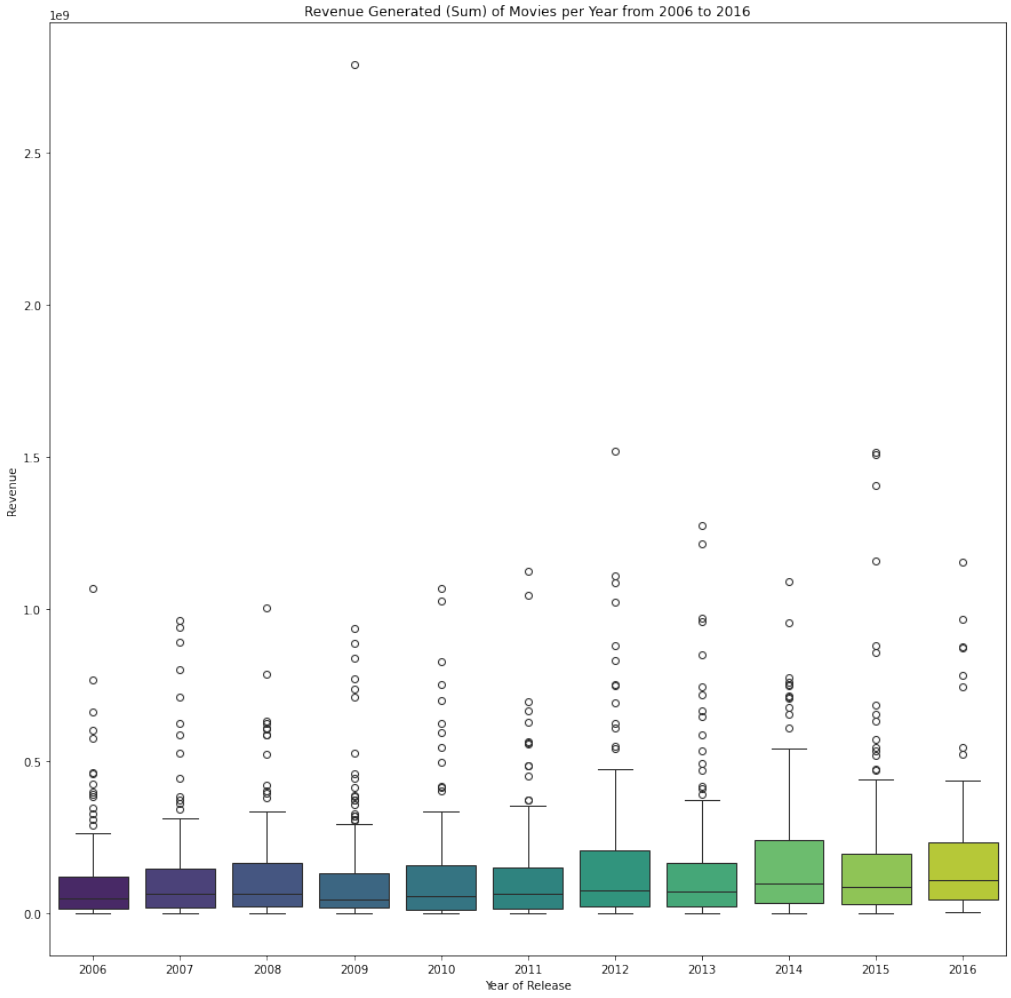

Figure 7

Movie runtime vs revenue

Figure 8 depicts the relationship between movie revenue and runtime, revealing a moderate positive correlation where longer movies tend to generate higher

revenue. (According to Investopedia, 2014). Several factors may contribute to this observation:

- Longer movies can tell bigger stories, appealing to viewers who enjoy deep plots and well-developed characters.

- Fans of certain types of movies might think longer ones are worth the money, so they buy more tickets.

- Movie theatres often charge more for longer films, especially for special formats like IMAX, which helps make more money even if the number of people watching stays the same.

- If people like longer movies, they'll talk about them, which keeps them popular and brings in more viewers.

- Big movie series often have lots of dedicated fans who want to see every instalment, making longer films in these series more popular and profitable.

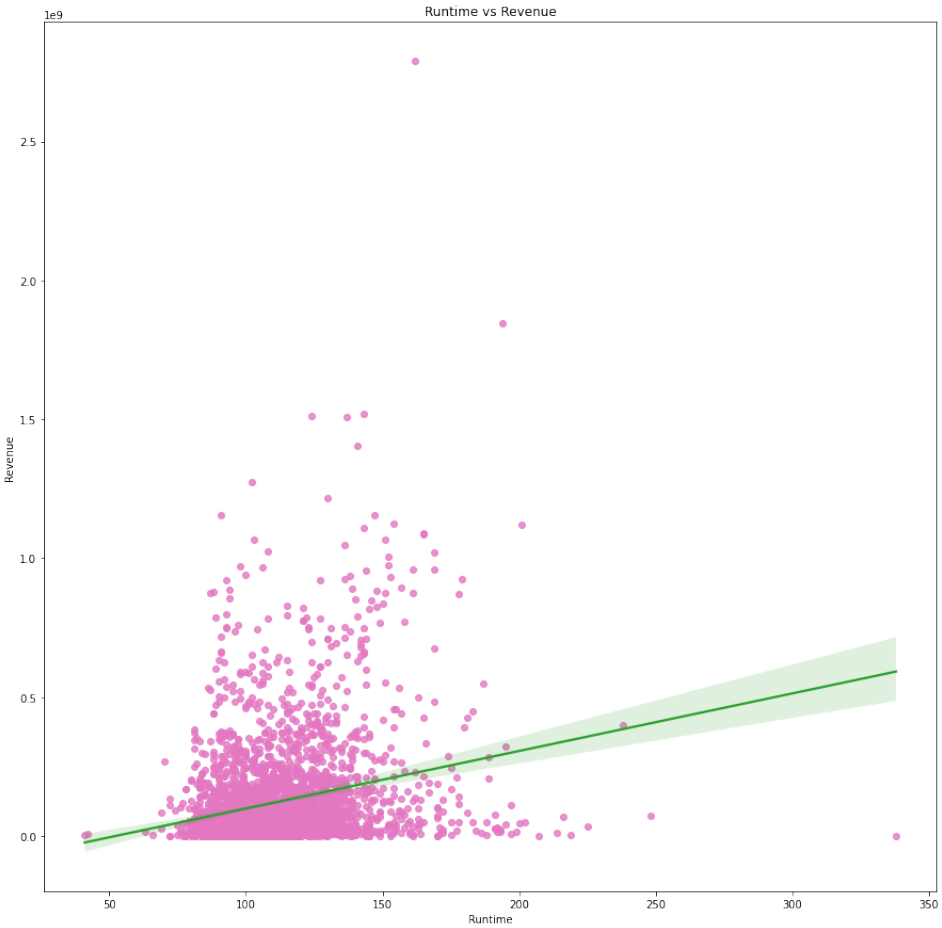

Figure 8

References

Investopedia. (2024). How Exactly Do Movies Make Money? Retrieved from https://www.investopedia.com/articles/investing/093015/how-exactly-do-movies-make-money.asp#:~:text=Key%20Takeaways-,While%20there's%20a%20lot%20of%20money%20to%20be%20made%20in,fickle%20public%20come%20into%20play.

ScreenCraft. (2024). 6 Essential Traits of a Great Drama Screenplay. [Online]. Available: https://screencraft.org/blog/6-essential-traits-of-a-great-drama-screenplay/[Accessed: [26-02-2024]].

Stone Soup. (n.d.). Why Animation Is Important. [Online]. Available: https://stonesoup.com/post/why-animation-is-important/[26-02-2024].

HyperionDev. (2021). Data Analysis - Preprocessing. [Educational notes]. Retrieved from Dropbox-NK23110009394.

THIS REPORT WAS WRITTEN BY : Nagitta Kasirye-Koikanyang