Python for Data Analytics Final Project: Pixar Films

GROUP 11

ALYSSA BRENNAN & KNOELLE GRASSI

Data Loading & Inspection

A Pixar films dataset was found on Kaggle. It has information about 28 Pixar films from 1995 to 2024. It has 28 rows and 15 columns.

Some columns in the dataset are:

- Ratings
- Release dates
- Box office totals
- Critic scores
- Audience scores

The data types include 11 integers, 1 float, and 3 strings.

There are missing values in the columns cinema_score and budget. These are due to being released around COVID and thus not getting the normal release process. The cinema_score missing values are given as N/A and the missing value in budget is given as 0.

Data Cleaning

The budget column had a single value that was 0 so a new column was made replacing that value with 'NaN'. Box_office_us_canada included a couple of outliers so a new column was made to identify these outliers. The rest of the data did not require any cleaning.

Descriptive Statistics

Some of the mean values found were:

- Run Time – 100.39 minutes
- Budget – $155,392,857.14
- US & Canada Box Office – $247,544,425.32
- Other Box Office – $608,664,319.14
- Rotten Tomatoes Score – 88.4
- Rotten Tomatoes Score Counts – 274
- Metacritic Score – 78.1
- Metacritic Score Counts – 43
- IMDB Score – 7.5
- IMDB Score Counts – 525,608

Some of the median values found were:

- Run Time – 100 minutes
- Budget – $175,000,000
- Worldwide Box Office - $533,878,228.50
- Rotten Tomatoes Score – 94.5
- Rotten Tomatoes Score Counts – 266
- Metacritic Score – 79.5
- Metacritic Score Counts – 41
- IMDB Score – 7.55
- IMDB Score Counts – 395,843

Some of the modes found were:

- Run Time – 100 minutes
- Budget – $200,000,000
- Rotten Tomatoes Score – 95, 97, 98
- The mode for the rest of the columns included every row

Some of the Standard Deviation found were:

- Run Time – 8.05 minutes
- Budget – $53,820,923.43
- US & Canada Box Office – $163,160,866.49
- Rotten Tomatoes Score – 13.4
- Rotten Tomatoes Score Counts – 84.2
- IMDB Score – 0.66
- IMDB Score Counts – 368,871.29

Looking at the histogram graphs of the data:

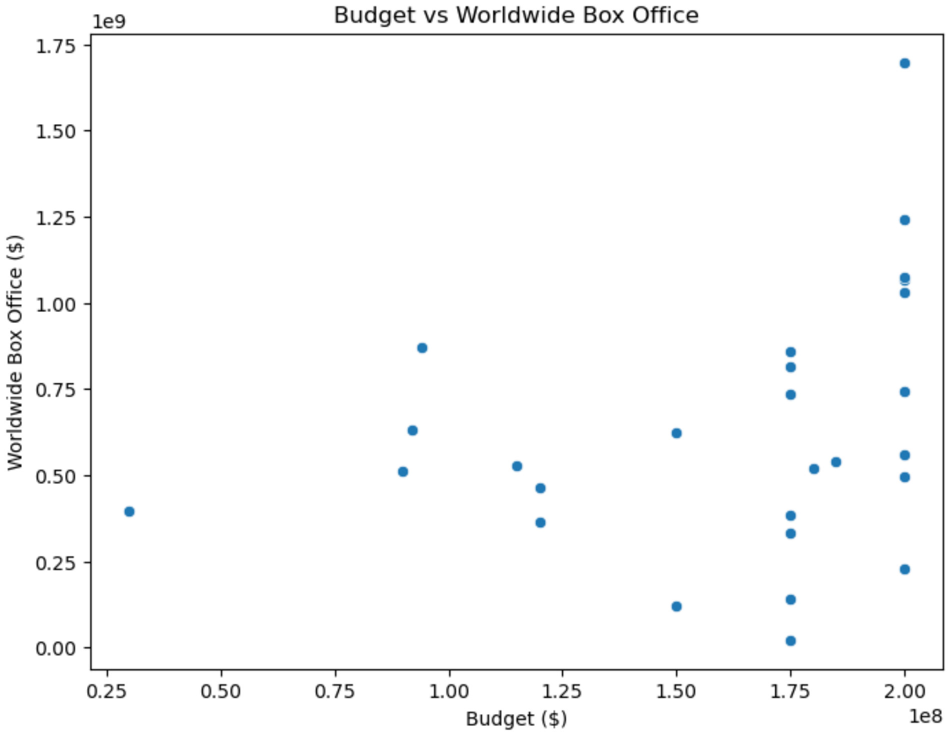
- A run time of 100 is most likely with a runtime of about 92 second most likely
- The higher the budget the higher the frequency of that budget
- The most frequent US & Canada Box office is 2e8 which is double the frequency then the next most likely box office
- The other box office usually has the smaller the box office the more frequent
- The worldwide box office frequency peaks at .38e9 and decreases after that
- Rotten Tomatoes is mostly around 90 and 100 with small peaks at 40 and around 60-80

- Metacritic peaks at 60 with a major drop before doubling after 70 to peak again at 75-85, it then halves before building to another peak at 95

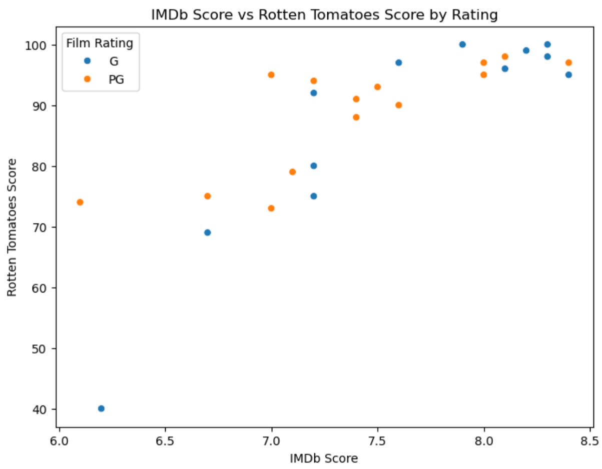
Looking at the box plots of the data:

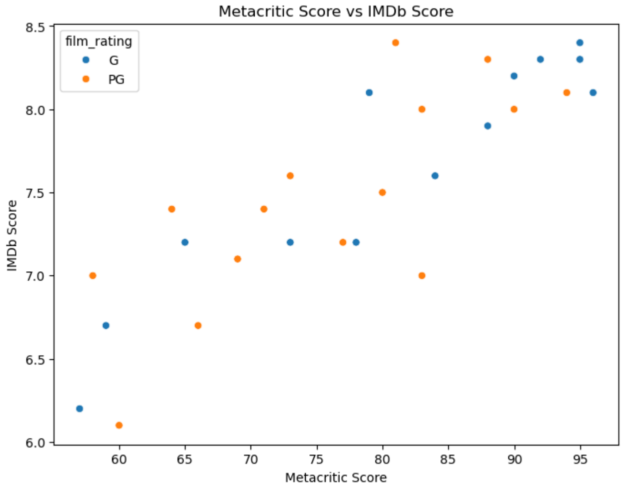
- 80 minutes is an outlier of the runtime and most of the data occurs between 95 minutes & about 105 minutes
- Most of the budget is between 1.25e8 and 2e8 with 25% occurring between 1.75e8 and 2e8
- US & Canada Box Office has the smallest distribution, other Box Office is next, and worldwide Box Office is the most spread out
- Rotten Tomatoes has a higher average than Metacritic with the highest 50% being around the top tail of Metacritic. Metacritic also has a wider range of scores, but Rotten Tomatoes has an outlier.
- IMDB is rated differently than the other two but seems to have spread in between the others and have the lowest scores.

Data Visualization



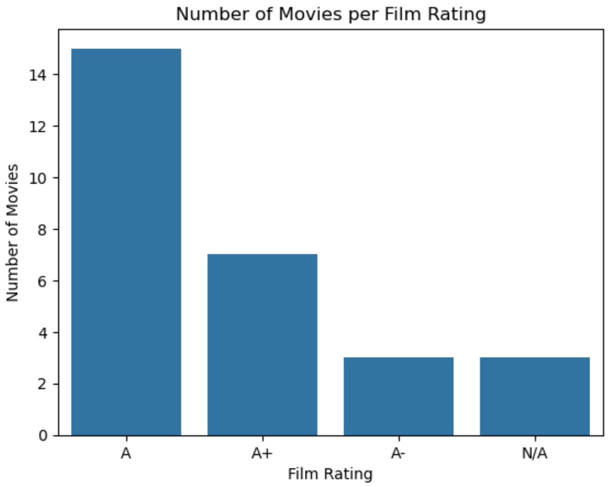
A scatterplot of budget vs worldwide box office showed that while most movies have a higher budget, there is not a clear relationship between budget and worldwide box office. Most of the data is between 1.5e8 and 2e8 of Budget.

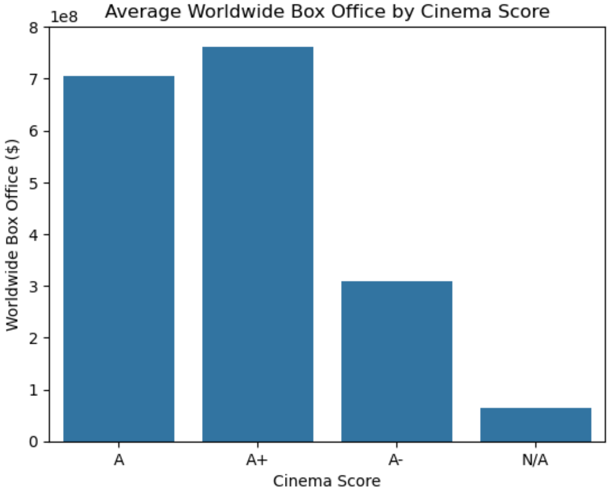




A scatterplot of IMDB score vs Rotten Tomatoes shows a higher IMDB score usually leads to a higher Rotten Tomatoes Score. There also seems to be no relationship between either score with the Film Rating. There are a couple of outliers at lower IMDB scores resulting in an average Rotten Tomatoes score.

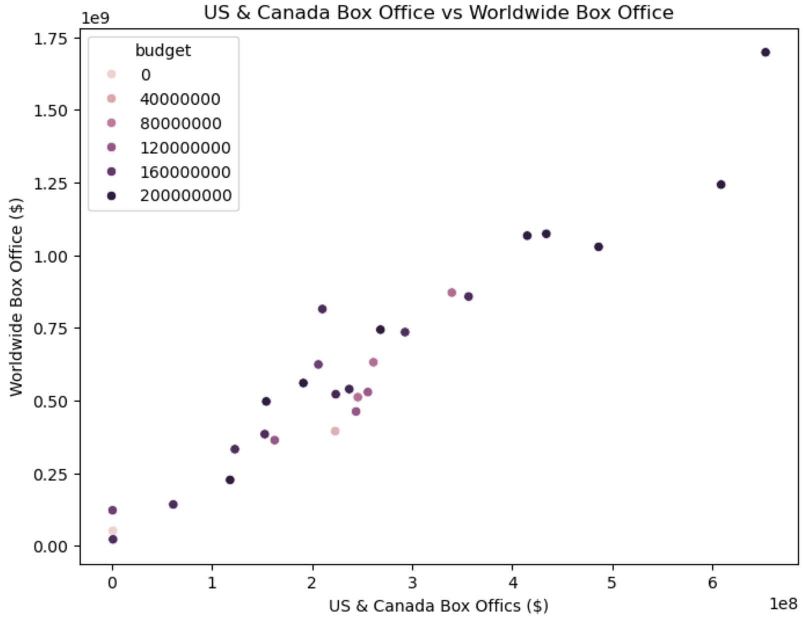
The scatterplot of Metacritic Score vs IMDB Score shows there is also a relationship between the two as a higher score in one result in a higher score in the other. There also seems to be a relationship that a lower score results in a lower score with the other.



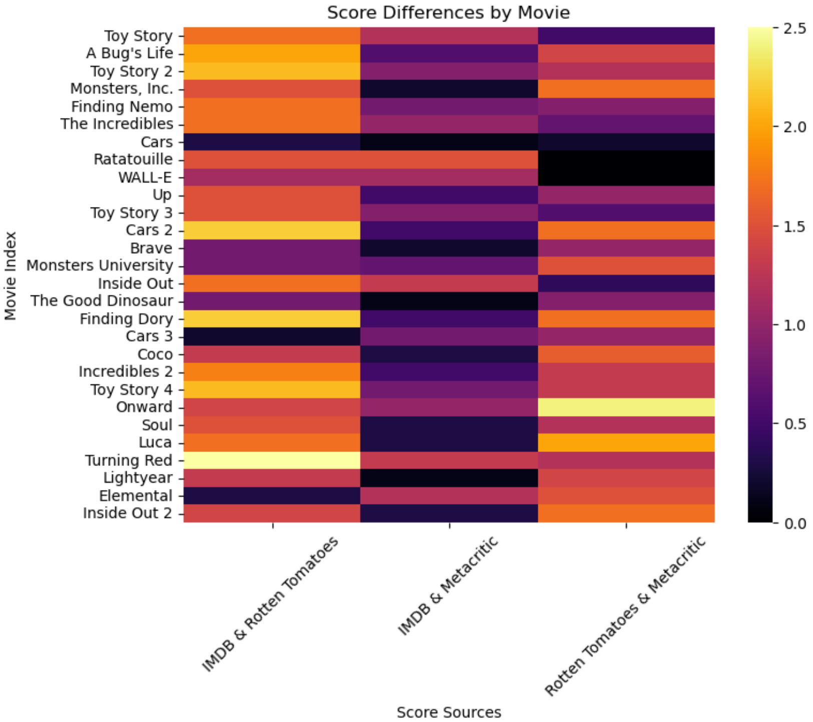


The bar graph of Number of Movies for each Film Rating shows all the movies released in cinemas receive a high score with most being A. This means all Pixar movies released in cinemas receive a high rating and are perceived as being pretty good.

The bar graph of Worldwide Box office vs Cinema Score shows the better the cinema score results in more in the worldwide box office. There is not much difference between A+ and A but there is a drop off to A-.

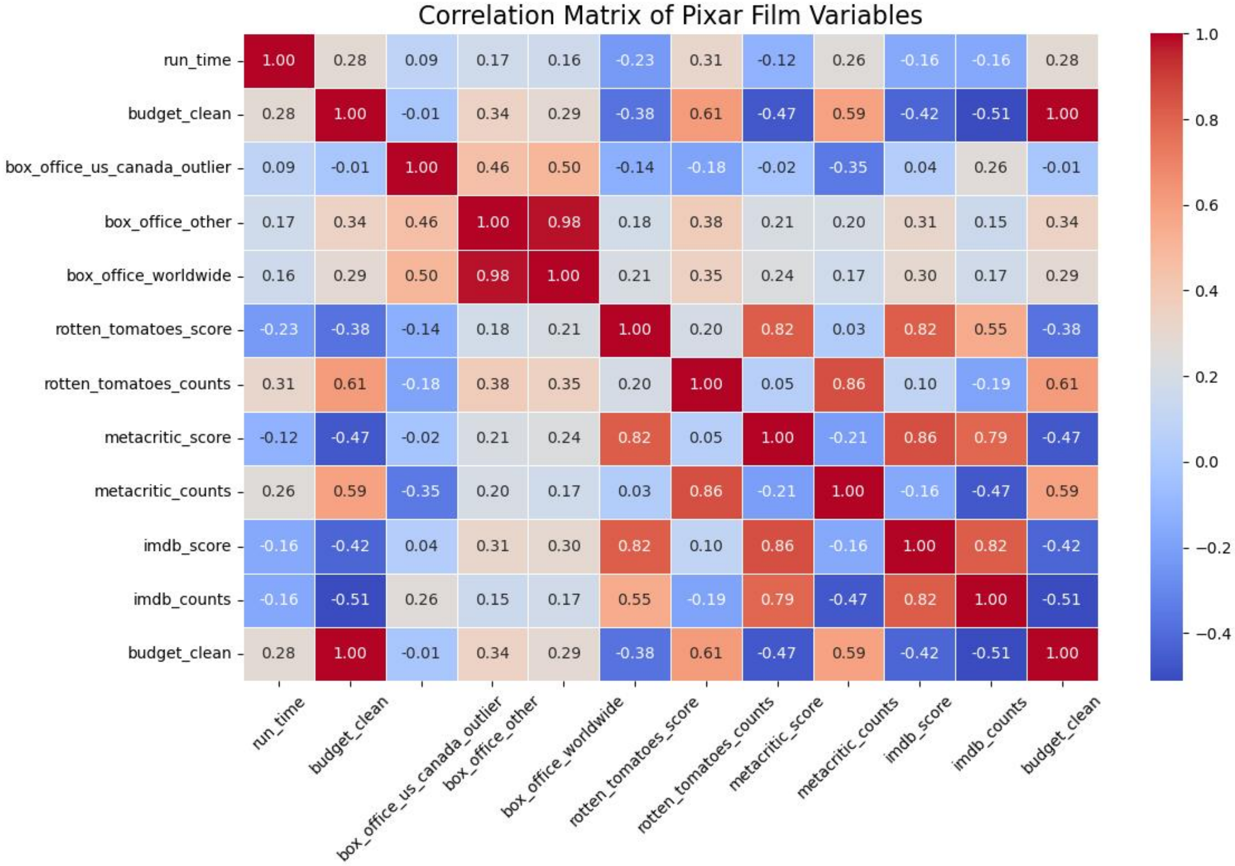


The scatterplot shows there is a relationship between the US & Canada Box Office and the Worldwide Box Office. A higher value in one results in a higher value of the other. There also seems to be some relationship between budget and the box office values. A higher budget results in a higher Box office, but the relationship doesn't seem as strong as the relationship between the box offices.



The heatmap of Score Differences of Critic Reviews shows that most of the time there is almost no difference between IMDB & Metacritic except for Ratatouille. There is usually a high difference between IMDB & Rotten Tomatoes except with Cars. The difference between Rotten Tomatoes & Metacritic is less than IMDB & Rotten Tomatoes but not as good as the difference between IMDB & Metacritic. This means IMDB & Metacritic are the closest related between the scores.

Correlation Analysis

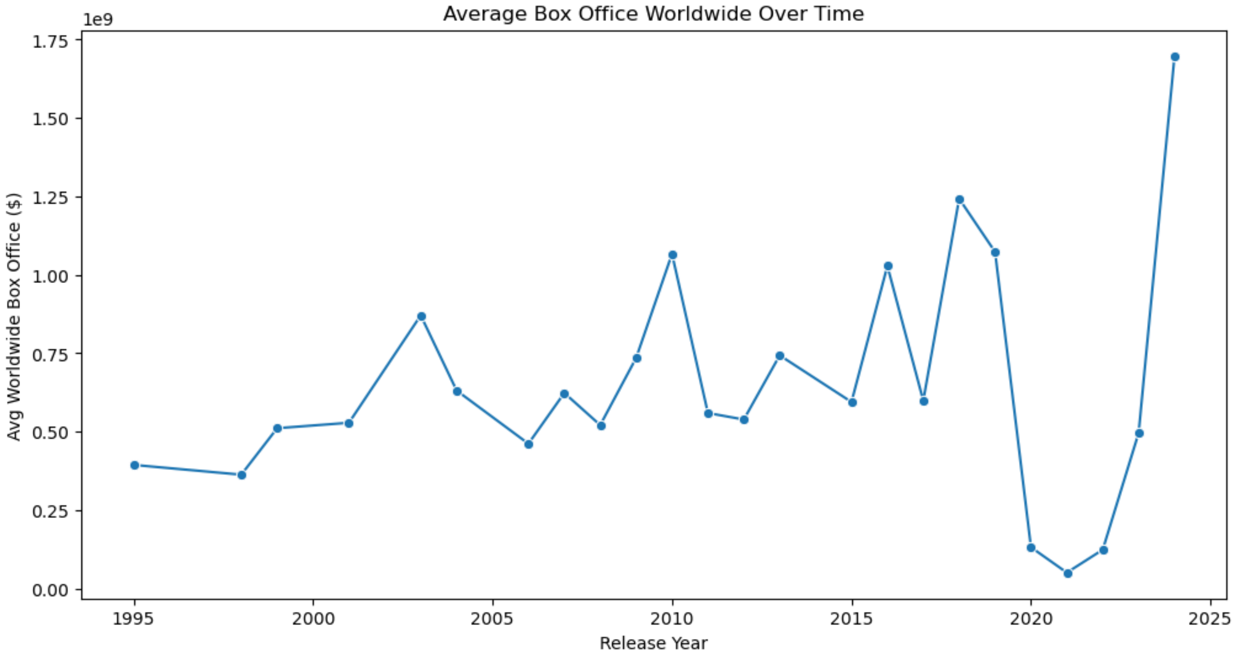


Box Office Worldwide and Box Office Other have the closet positive relationship and nearly perfectly correlated. This makes sense as the worldwide total is largely influenced by the international box office (box office excluding Canada and US). Surprisingly Box Office US & Canada is not more closely related to either of the other Box Office, and while they do have a positive relationship it's not as strong as the others are to each other. This shows that domestic success does not always mirror international success.

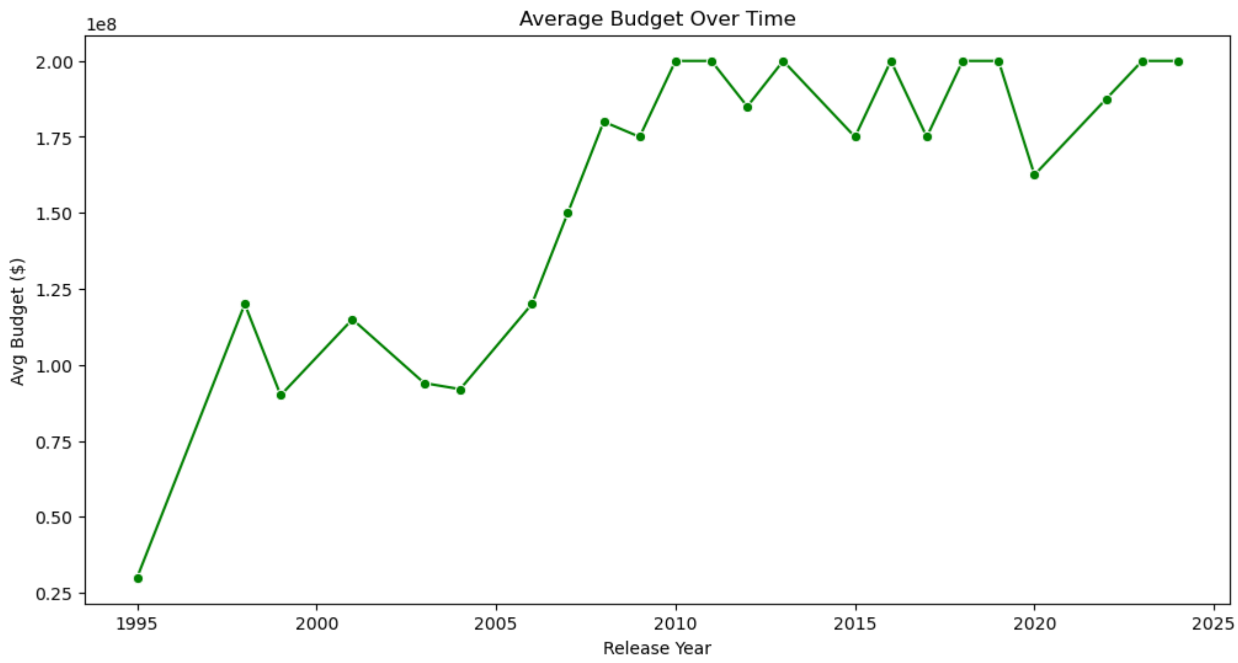
Rotten Tomatoes Score and Metacritic Score also have strong correlation found earlier with the visualizations. Rotten Tomatoes also has a strong relationship with IMDB score. Metacritic Score and IMDB Score also have a strong relationship. This means reviewers usually agree across the three platforms. However, they all have a low to moderate correlation with box office which means even through the films review good they do not always make money. They also all have a negative correlation with the budget. This shows that a higher budget does not mean good reviews. However, the rotten tomatoes score counts and Metacritic score counts are the best columns correlated with budget. This means the films with higher budgets attract more reviewers. IMDB Score is also the only platform to have a strong correlation with its number of scores. IMDB Score Count also has a strong correlation with Metacritic Score and a moderate correlation with Rotten Tomatoes Score. This could reflect how more popular films tend to be rated higher on the platforms.

Run time has low or negative correlation with every other column. This means it does not strongly predict budget, box office earnings, or critical reception.

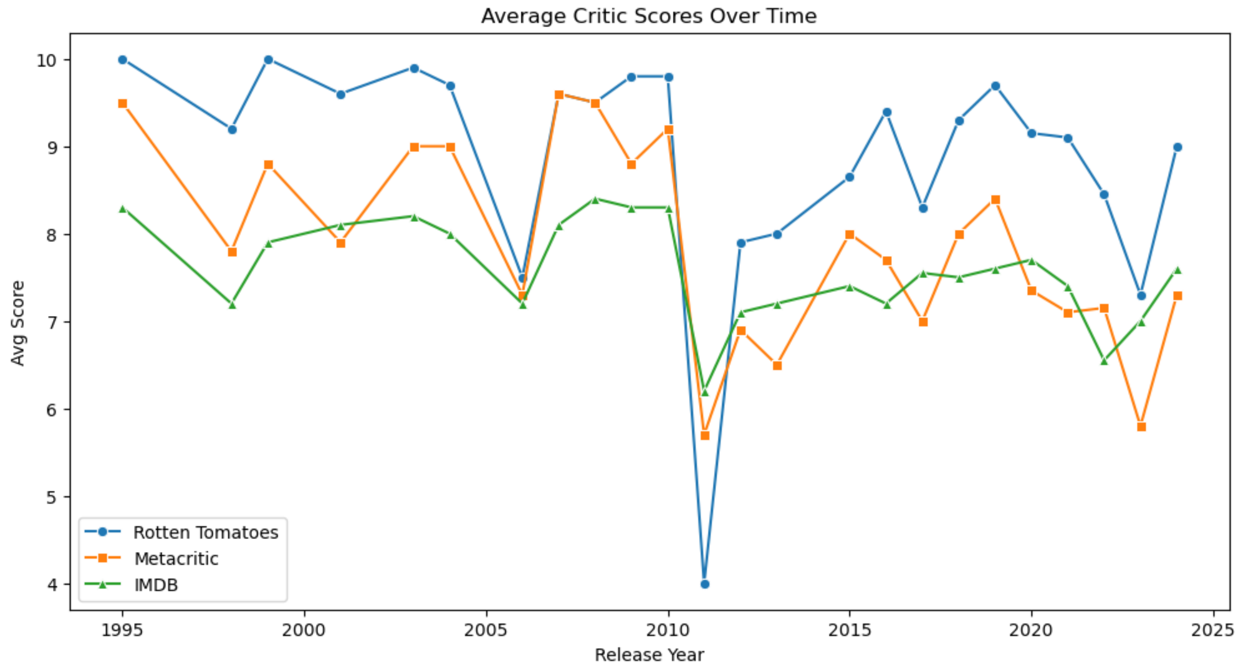
Time Series Analysis



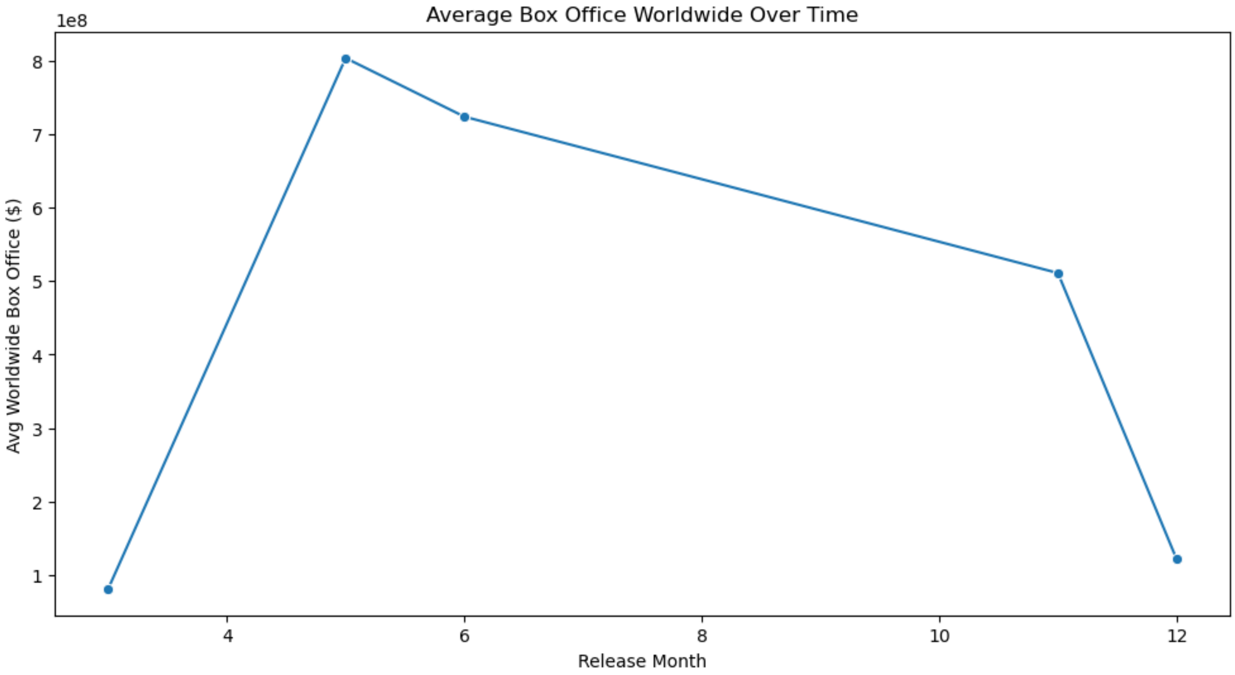
The box office has been increasing since 1995 with small dips such as in 2005 and 2020 before continuing to go up. 2025 has the highest average box office worldwide while 2021 has the lowest (COVID). The more recent movies, after 2020, have made more money than most of the older movies.



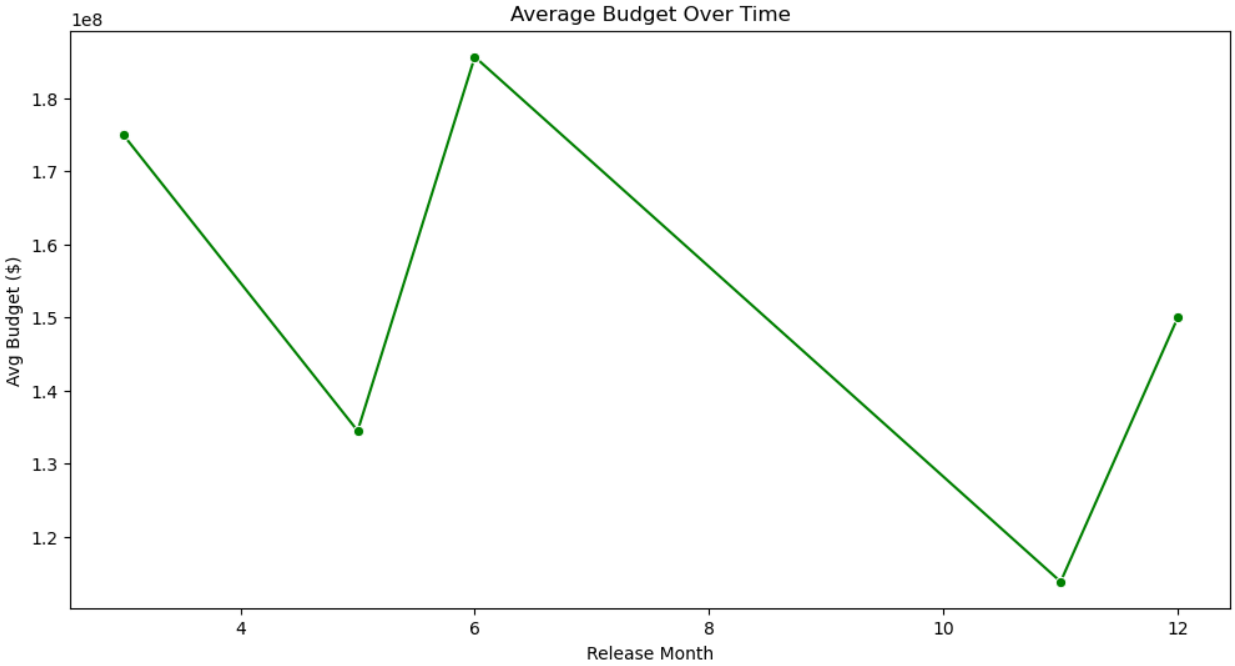
The budget has been increasing since 1995 and peaks every couple year since 2010 before decreasing again and increasing again after. The more recent the movie the more budget they have.



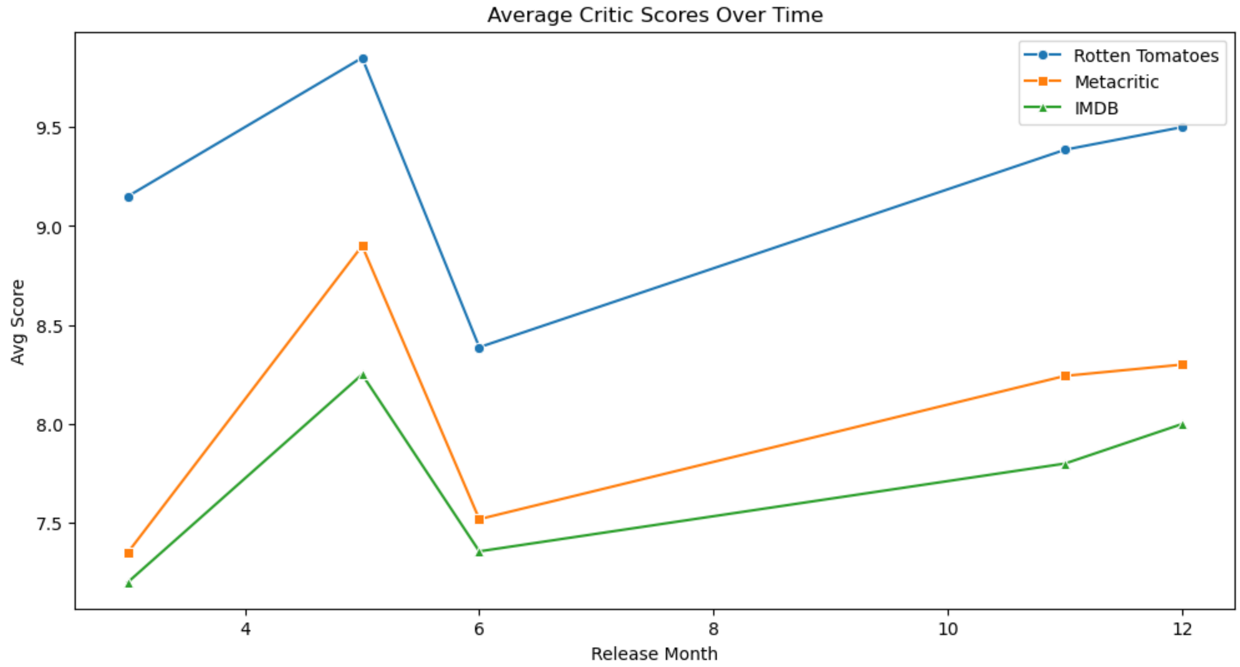
The average critic score has been decreasing across time with the lowest occurring since about 2011. The more recent movies have not been as critically successful as the historical movies.



The movies released in May have the highest worldwide box office. After May they decrease slightly to June and continue declining until November. January and December have the lowest box office. After January it starts increasing again until May. The movies released over the summer make the most money while the movies released over the winter make the least.



The movies released in June have the highest budget followed by January. November has the smallest budget. This means movies released in June take the most money to make while movies in November are the least expensive. There does not seem to be much of a seasonal trend to the budget.



The movies released in May have the best critic score. There is a drop until June when it starts increasing before dropping in January before increasing to May. January and June have the lowest critic score. There does not seem to be a seasonal trend for critic scores.

Key Takeaways

Pixar movies generally have high critical reception with Rotten Tomatoes, Metacritic, and IMDB scores all showing strong correlation. Budget and worldwide box office have a weak positive relationship, but higher budgets do not guarantee critical success. The strongest box office correlations were between international and worldwide totals, meaning domestic performance does not always match global success. Critical scores have slightly declined over time, especially after 2011, which budgets and box office earnings have generally increased. May and June are peak months for box office returns, especially May. May is also the highest month for critical scores. Movies released in the summer perform better financially while winter releases underperform. There is no strong correlation between run time and any other feature.

We had challenges with some data fields having missing or placeholder values. This was addressed by converting them to NaN and using new columns to handle outliers and visualize trends more accurately. The sample size was small so it was difficult to draw conclusions for relationships as some trends could have been exaggerated due to outliers.

The critic scores being different scales was also a challenge that was solved by putting them all on the same scale.

Future analysis could include other movies to compare to such as DreamWorks and Illumination or include Pixar short films and series to strengthen trend reliability. Include data from social media or other review platforms like Letterbox to analyze more reviews. Include the theme or plot of movies to see if they play any part in critical and box office success. Include marketing budget and distribution details to see how those contribute to critical and box office success. Include if something else was going on to influence any of the data such as COVID in 2020/2021.

Task Summary Table

Data Loading & Inspection	Knoelle Grassi	1 Hour
Data Cleaning	Alyssa Brennan	1 Hour
Descriptive Statistics	Alyssa Brennan	2 Hours
Data Visualization	Knoelle Grassi	2 Hours
Correlation Analysis	Knoelle Grassi	1 Hour
Time Series Analysis	Alyssa Brennan	1.5 Hours
Report Writing	Knoelle Grassi	3 Hours
Presentation Preparation	Knoelle Grassi	2 Hour
	Alyssa Brennan	1 Hour