Data Process

- What software or systems did you end up using?

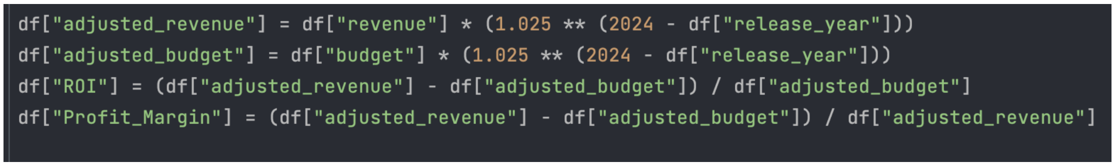
(Ben)

We decided that using an Azure data factory was the best solution for analyzing our dataset. To utilize our data factory, we had to create a storage account where we uploaded our IMBD movie dataset to a container. Before we could upload this dataset, we had to clean up our data. Using Python, we simply removed unnecessary or irrelevant columns from our dataset. We also derived new columns by performing various calculations to output a new dataset that we could use for this project. Within our data factory, we could create our own data flow, where we first read in our modified csv dataset as a source. From there, we could query our data by adding filters where we could gather desired insights within our dataset. Finally, after querying our results, the last thing to do is visualize our data for more efficient analysis. In doing so, we utilized tools such as Excel and PowerBI to create visuals that would shape our output into something that we could easily interpret.

- What steps did you take?
    - Did you create any new columns (calculated from original columns)

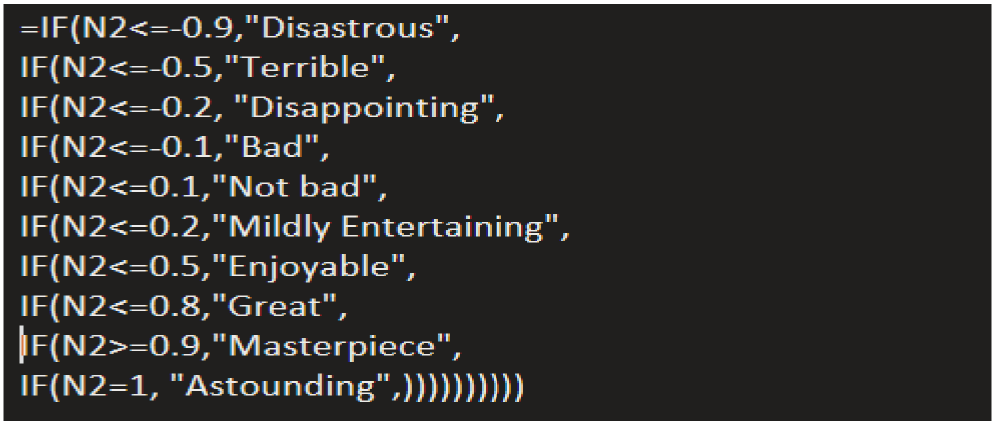
(Ben)

Our original dataset involved several financial metrics such as revenue and budget for producing respective movies. In the early stages of our project, we understood that we would need to account for the difference in release years when analyzing these metrics. Due to inflation, a movie that generates 100 million in revenue in the 90s is much more successful than one producing the same number in 2024. Therefore, we implemented an annual inflation factor of 2.5% to account for average yearly inflation. As you can see in the code below, we took that inflation factor to the power of the difference between the current year (2024), and each movie's respective release year. Though inflation has not been consistent every single year, using this inflation factor gives us a much better way to compare movies with large disparities in release year. We also calculated several profitability metrics such as Profit Margin and ROI to analyze the financial performance of each movie.



```
df["adjusted_revenue"] = df["revenue"] * (1.025 ** (2024 - df["release_year"]))
df["adjusted_budget"] = df["budget"] * (1.025 ** (2024 - df["release_year"]))
df["ROI"] = (df["adjusted_revenue"] - df["adjusted_budget"]) / df["adjusted_budget"]
df["Profit_Margin"] = (df["adjusted_revenue"] - df["adjusted_budget"]) / df["adjusted_revenue"]
```

We performed sentimental analysis by evaluating keywords from the reviews, movie and taglines to analyze overall sentiment. We categorized movies into different sentiments categories.



```
=IF(N2<=-0.9,"Disastrous",
IF(N2<=-0.5,"Terrible",
IF(N2<=-0.2, "Disappointing",
IF(N2<=-0.1,"Bad",
IF(N2<=0.1,"Not bad",
IF(N2<=0.2,"Mildly Entertaining",
IF(N2<=0.5,"Enjoyable",
IF(N2<=0.8,"Great",
IF(N2>=0.9,"Masterpiece",
IF(N2=1, "Astounding",))))))))))
```

- Did you filter your data

We filtered our dataset for sentimental analysis to include only movies with a minimum of 3000 reviews. This was established to ensure that sentimental analysis enhances the reliability of our results and reflects significant audience opinions.

- Did you aggregate your data

- Where did your data go? Share the journey of where it flowed

Please include screenshots of your database work! They help the instructor and class understand better what steps you took!

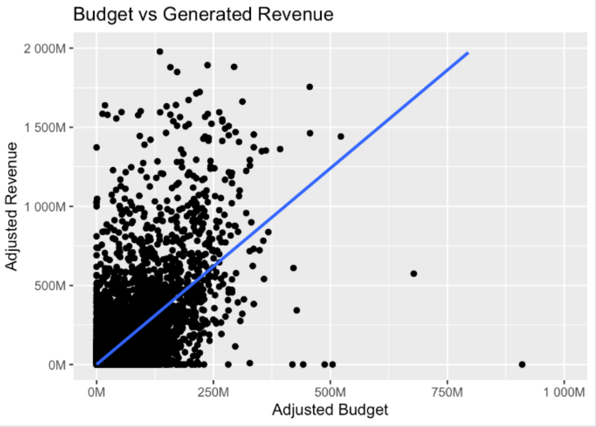
Assumptions Made

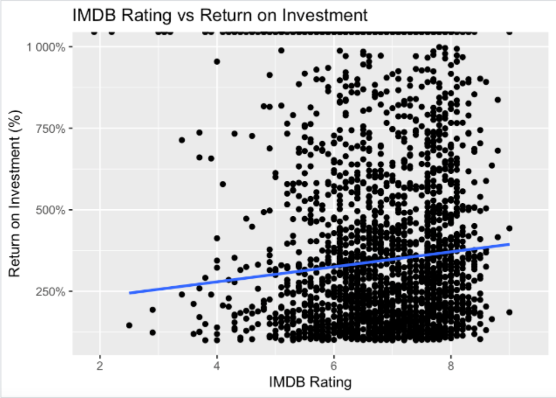
This section should include assumptions your team made in order to do this analysis. Consider the assumptions we make when doing a statistical analysis. Some examples may include:

- Assumptions about how the data was collected (was it a simple random sample? Is it possible the data may be biased due to collection methods?)
- Variance assumptions
- Was the sample size large enough?

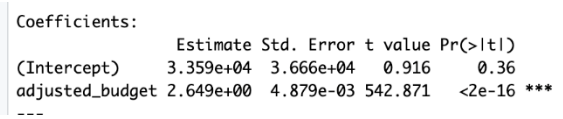
Visualizations

Linear Regression Plots:

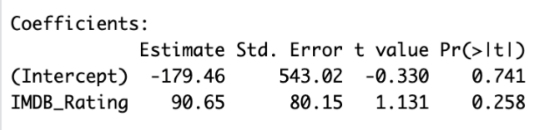




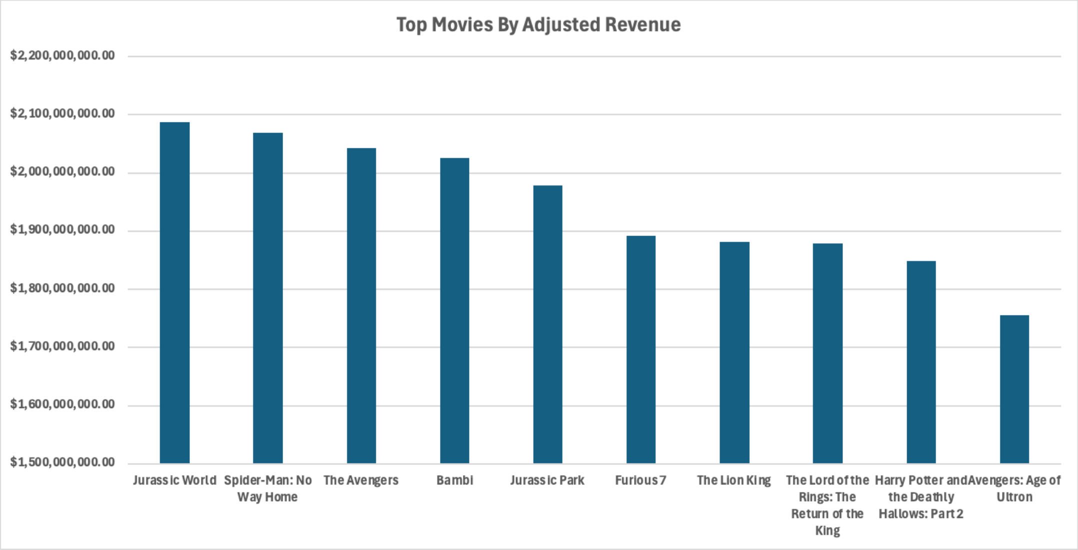
Model Fits for Respective Plots Below:

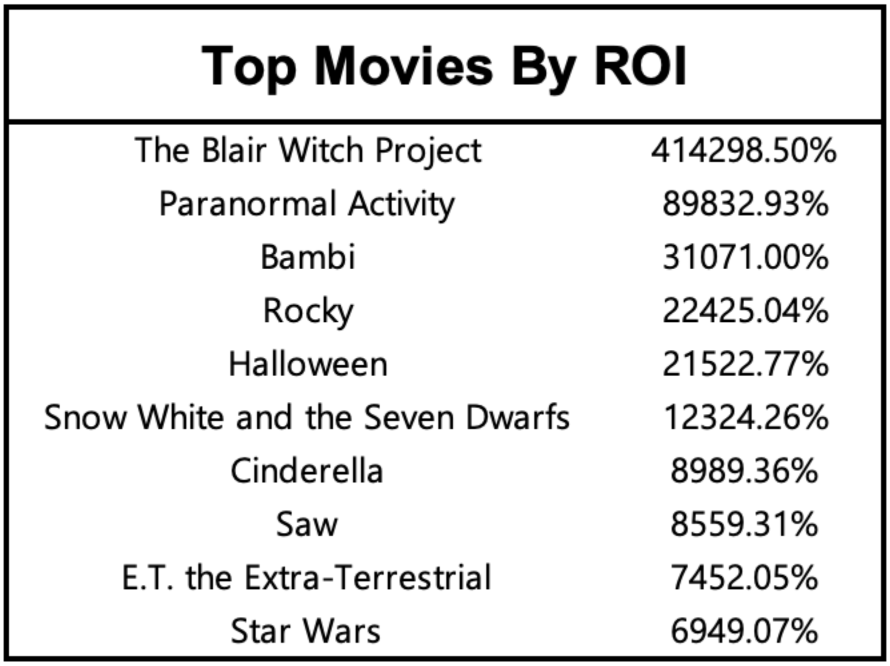


```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.359e+04  3.666e+04   0.916     0.36
adjusted_budget 2.649e+00  4.879e-03 542.871   <2e-16 ***
---
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -179.46     543.02  -0.330    0.741
IMDB_Rating    90.65      80.15   1.131    0.258
```



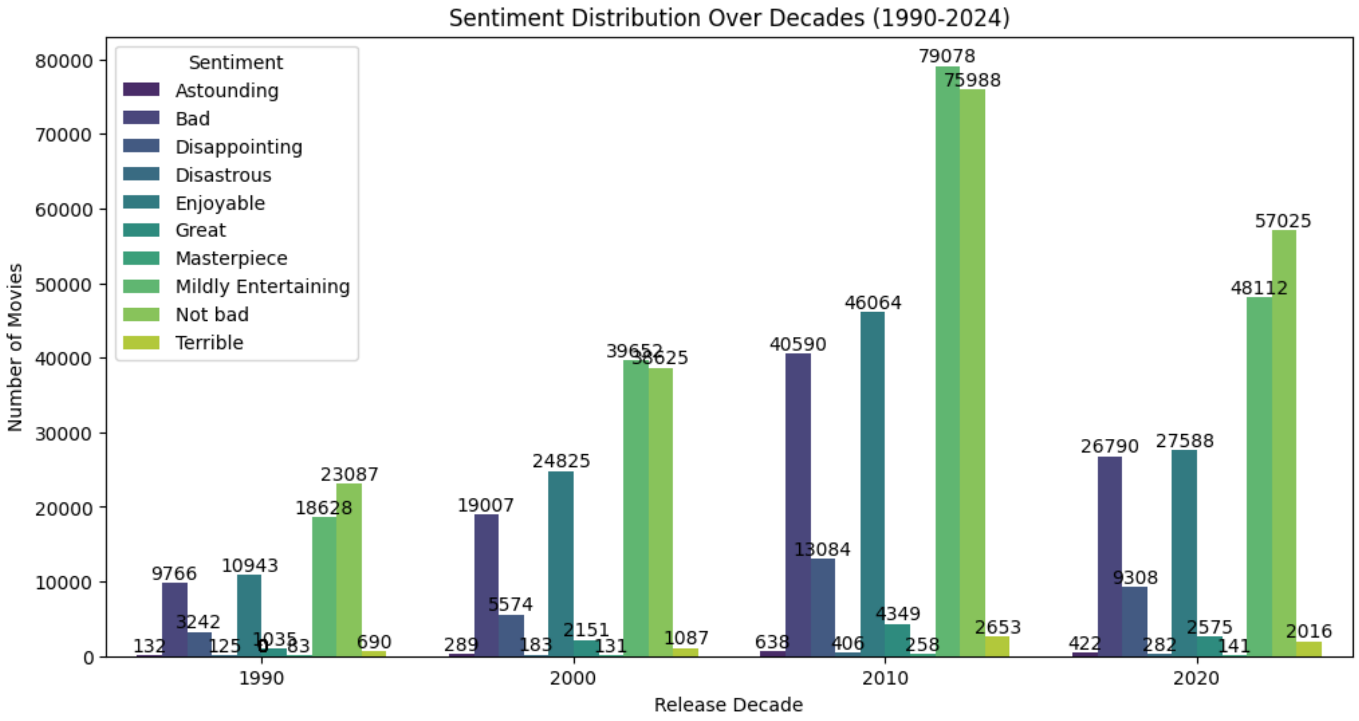


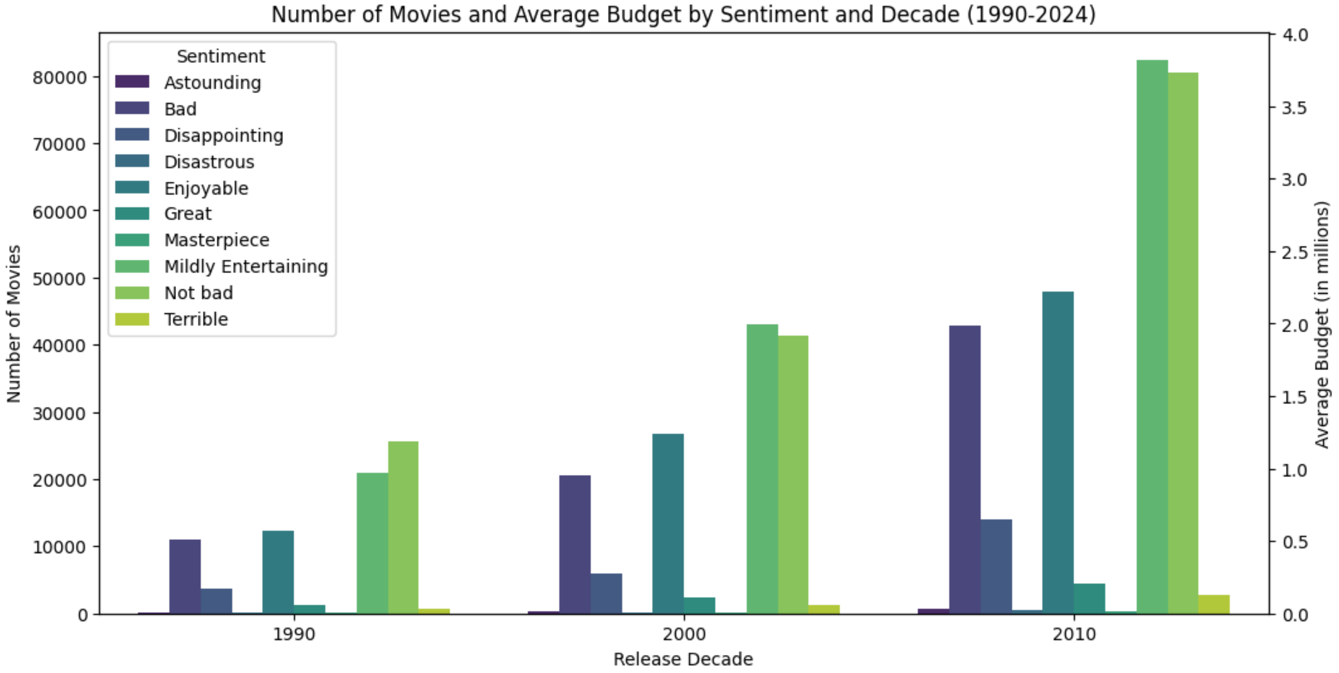
Top Movies By ROI

The Blair Witch Project	414298.50%
Paranormal Activity	89832.93%
Bambi	31071.00%
Rocky	22425.04%
Halloween	21522.77%
Snow White and the Seven Dwarfs	12324.26%
Cinderella	8989.36%
Saw	8559.31%
E.T. the Extra-Terrestrial	7452.05%
Star Wars	6949.07%

Table: The listed movies below are the highest budget (adjusted) corresponding to the sentimental score from each category.

Index	Title	Sentiment	Profit Margin	Adjusted budget	Adjusted revenue
1	Sing	Astounding	0.881731765	91380217.31	772652243.2
2	Pirates of the Caribbean: At World's End	Bad	0.687825182	456485478.4	1462275149
3	Avengers: Endgame	Disappointing	0.872857143	402781323.8	3167942996
4	Eragon	Disastrous	0.598857715	155965871.8	388804366.2
5	Titanic	Enjoyable	0.911667112	389560003.7	4410135473
6	Incredibles 2	Great	0.839073755	231938683.6	1441273195
7	Avatar: The Way of Water	Mildly Entertaining	0.801745526	483287500	2437712951
8	Pirates of the Caribbean: On Stranger Tides	Not bad	0.637568138	522455686	1441528026
9	Night at the Museum: Battle of the Smithsonian	Terrible	0.636897217	217244725	598300908.6





Recommendations

What recommendations or data-driven insights would you make based off of your results?

(Ben)

In this project we tried to take a look at how some potential explanatory variables could impact the financial performance of various movies. It came as no surprise to us that there was a positive relationship between movie's budgets and their generated box office revenue. Using our fitted model, we can estimate that on average, a movie will generate an additional $2.65 in revenue for each additional dollar spent in the budget. There is a very wide variance within this relationship, so trying to project revenue solely based on budget isn't going to be very accurate. We also used regression to explain the relationship between the IMDB rating and the movie's ROI. We also noticed a positive relationship between these variables. On average, for each additional 1 unit increase in the IMBD rating, the movie's ROI increased by roughly 90%. Though these models accurately tell us about the relationship between these variables, they are skewed by significant outliers within our data. Some movies (such as Blair Witch Project and Paranormal Activity) have outrageous ROIs that contribute to skewing of the model fit.

- IMDb-TMDB rating comparison across different fields – Michael
-
- – Prajakta

How do these findings impact your company?

Understanding sentiment can improve budget allocation, focusing on the projects which can be more successful at the box office and further helping the production houses to decide theatre or online platforms in which the movie will perform well. These findings can guide us to strategic decisions on marketing and project development by aligning the audience preference of movies which attract them to theatres.

How do them impact regular people like you and me?

Everyday evolving language and themes in movies it's a challenge to find a good child friendly weekend options. By using sentimental analysis, viewers can make more informed choices about movies which to watch, helping them find films aligned to their preferences. Additionally, this analysis helps to find new trends allowing an audience to better understand the latest trending content.

What should we do with these insights?

Please only submit 1 file per group

Notes:

- Include any piece of code relevant in the Appendix of your document
- Make sure to add enough detail so anyone that reads your report is able to reproduce your analysis