GROUP 1 - IB3K50

Analysis of the Movie Dataset

2049594, 2026661, 2016942, 2036682
1931258, 1924190, 1721901, 2053845

GROUP 1 - IB3K50

Movie Data **Summary**

Large amount of data scattered over time

- 4500+ movies produced in the span of more than a century
- This makes the analysis challenging as there is the possibility of deducing historical trends during the 20th and early 21st centuries

Variables used for regression analysis

- Revenue, budget, vote average, vote count, runtime, genres and months
- Genres provided as non-numerical data, we used R to analyze the number of genres per movie and their impact on popularity.

How to consider profit in our analysis?

- We have taken the difference between revenue and budget
- Production companies spending more on movies does not strictly warrant higher popularity

Key insights from Regression

- The number of genres in a movie impacts its popularity positively because it appeals to a wider range of audience
- Higher ratings and higher reviews also positively affect the popularity of the movie
- Movies released in August and December tend to have higher popularity

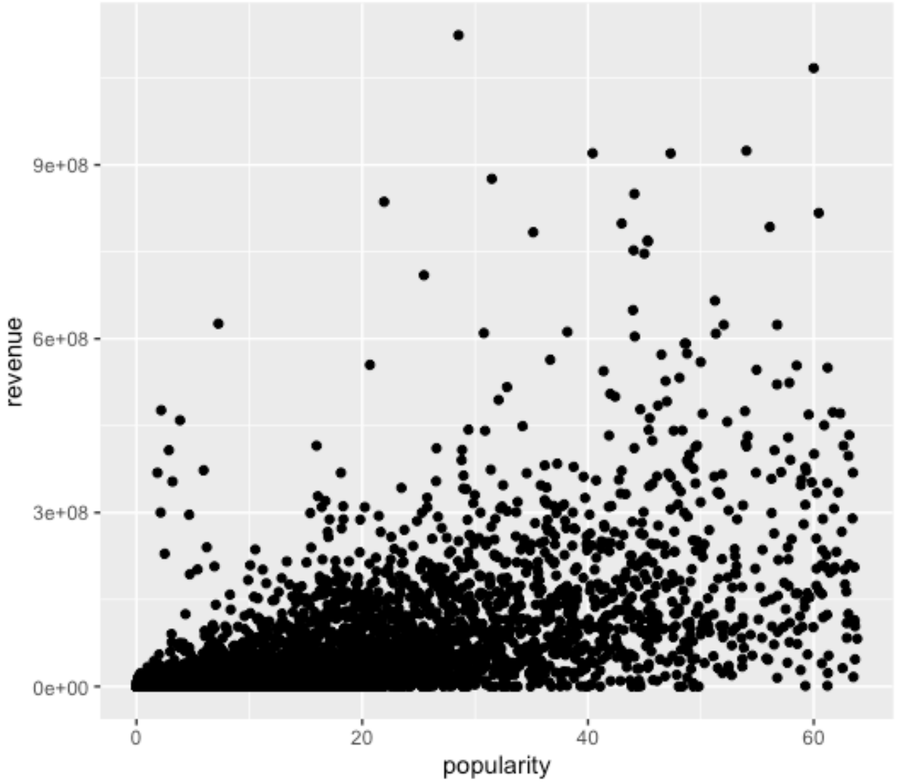
GROUP 1 - IB3K50

Regression Analysis

Regression Analysis

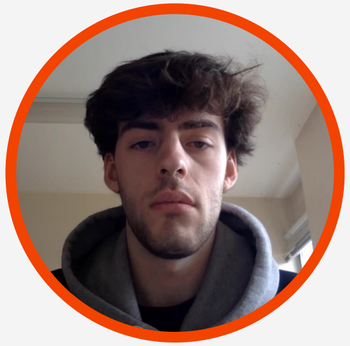
```
Call:
lm(formula = popularity ~ budget + vote_average + vote_count +
    num_genre + runtime + month_2 + month_3 + month_4 + month_5 +
    month_6 + month_7 + month_8 + month_9 + month_10 + month_11 +
    month_12, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-94.933  -4.807  -1.123   4.119  35.665

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.566e+00  8.894e-01  -2.885  0.00393 **
budget       6.517e-08  4.691e-09  13.891  < 2e-16 ***
vote_average 1.379e+00  1.190e-01  11.587  < 2e-16 ***
vote_count   1.539e-02  2.253e-04  68.323  < 2e-16 ***
num_genre    3.489e-01  1.165e-01   2.994  0.00277 **
runtime      5.551e-03  6.307e-03   0.880  0.37883
month_2      4.836e-01  6.462e-01   0.748  0.45427
month_3      8.030e-01  6.298e-01   1.275  0.20236
month_4     -1.998e-01  6.368e-01  -0.314  0.75367
month_5     -1.046e+00  6.451e-01  -1.621  0.10509
month_6     -2.892e-01  6.344e-01  -0.456  0.64849
month_7      8.380e-01  6.429e-01   1.303  0.19248
month_8      9.851e-01  6.090e-01   1.618  0.10582
month_9      2.250e-01  5.631e-01   0.400  0.68943
month_10     1.488e-01  5.918e-01   0.252  0.80141
month_11     5.016e-01  6.624e-01   0.757  0.44894
month_12     1.219e+00  6.111e-01   1.995  0.04610 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.454 on 4509 degrees of freedom
Multiple R-squared:  0.6888,    Adjusted R-squared:  0.6877
F-statistic: 623.9 on 16 and 4509 DF,  p-value: < 2.2e-16
```

Relationship between Revenue and Popularity

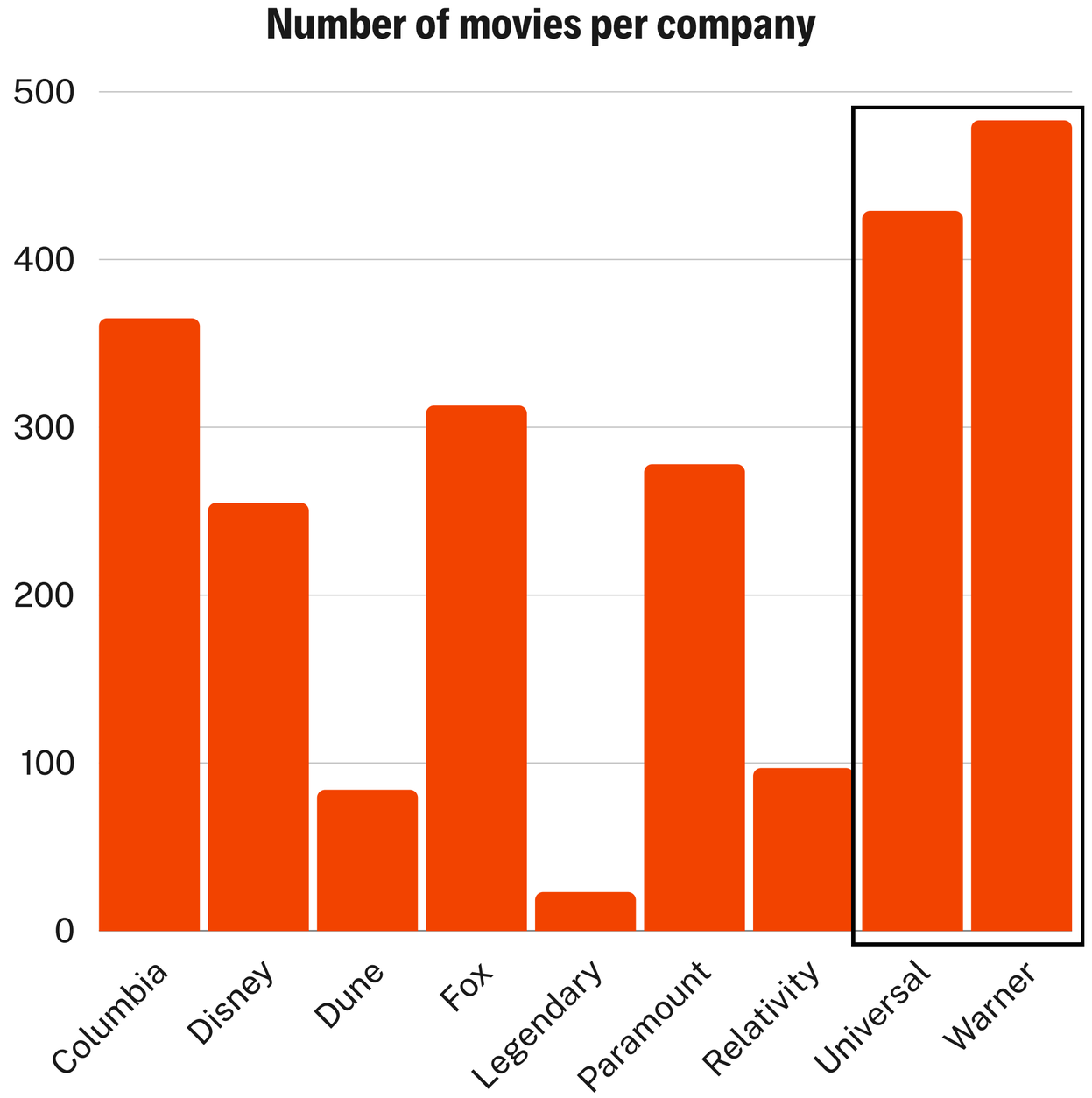


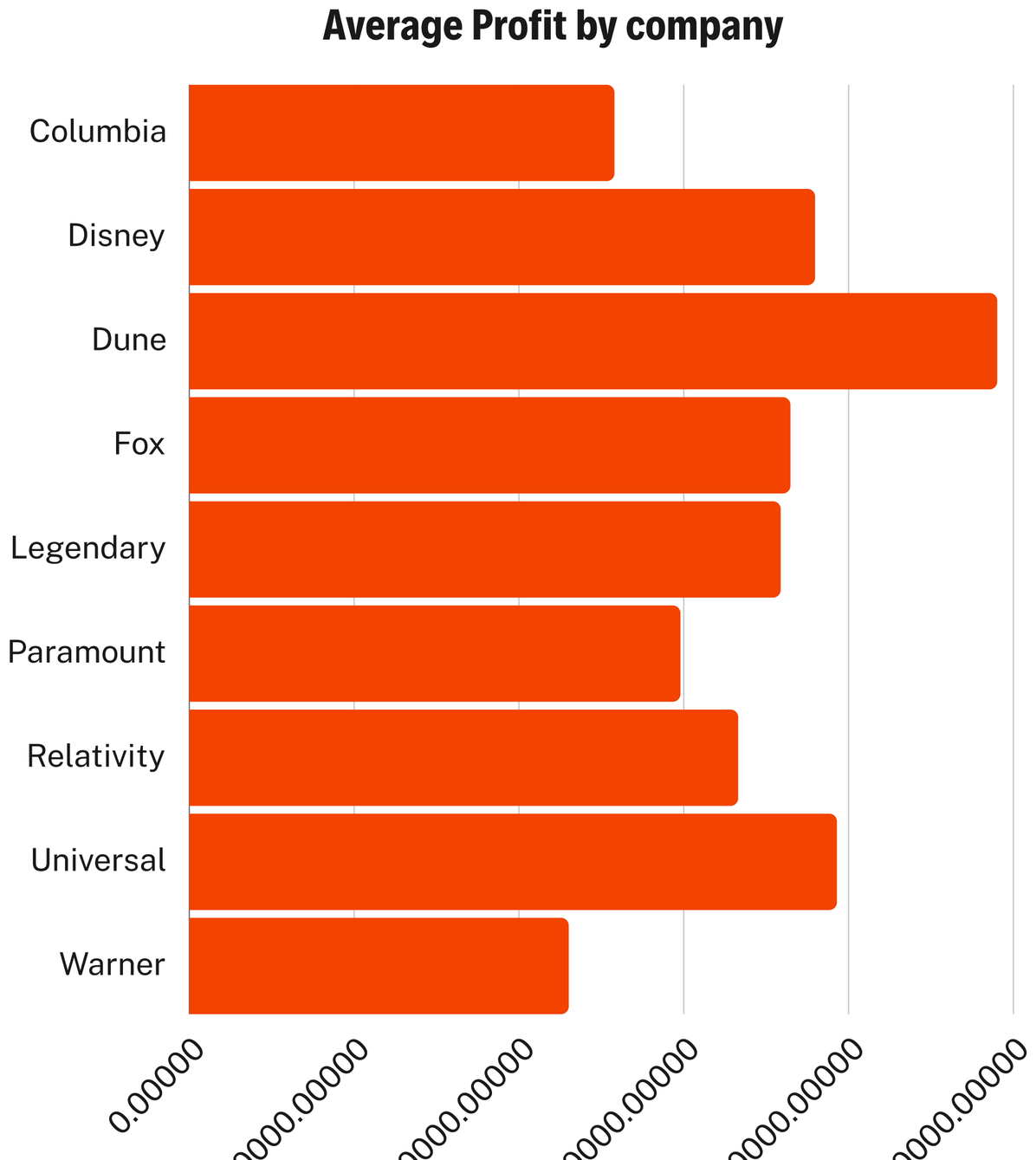
AI FOR BUSINESS

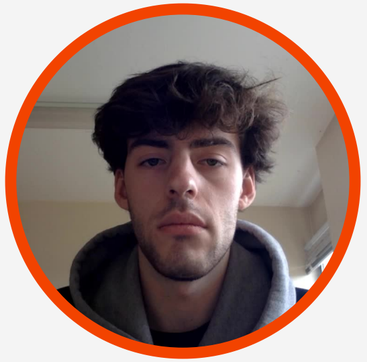
Overview of the Top 10 Companies



1. We have chosen to group Marvel with Disney as they are part of Disney Group
2. Significant gaps on the number of films produced by companies between 1916 and 2017
3. 3 types of companies: large (over 400), medium (between 200 and 400), and small (less than 200) film producers.
4. Warner and Universal have similar amount of movies (483 against 429)







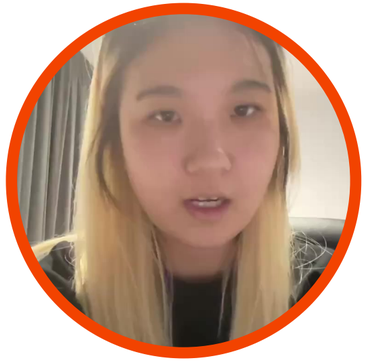
Focus on **Warner & Universal**

1. Releasing large number of films **does not necessarily mean** high profit for the company.

2. While Warner and Universal produced about 400 films between 1916 and 2017, Warner **only earned about half** of Universal's average profit.

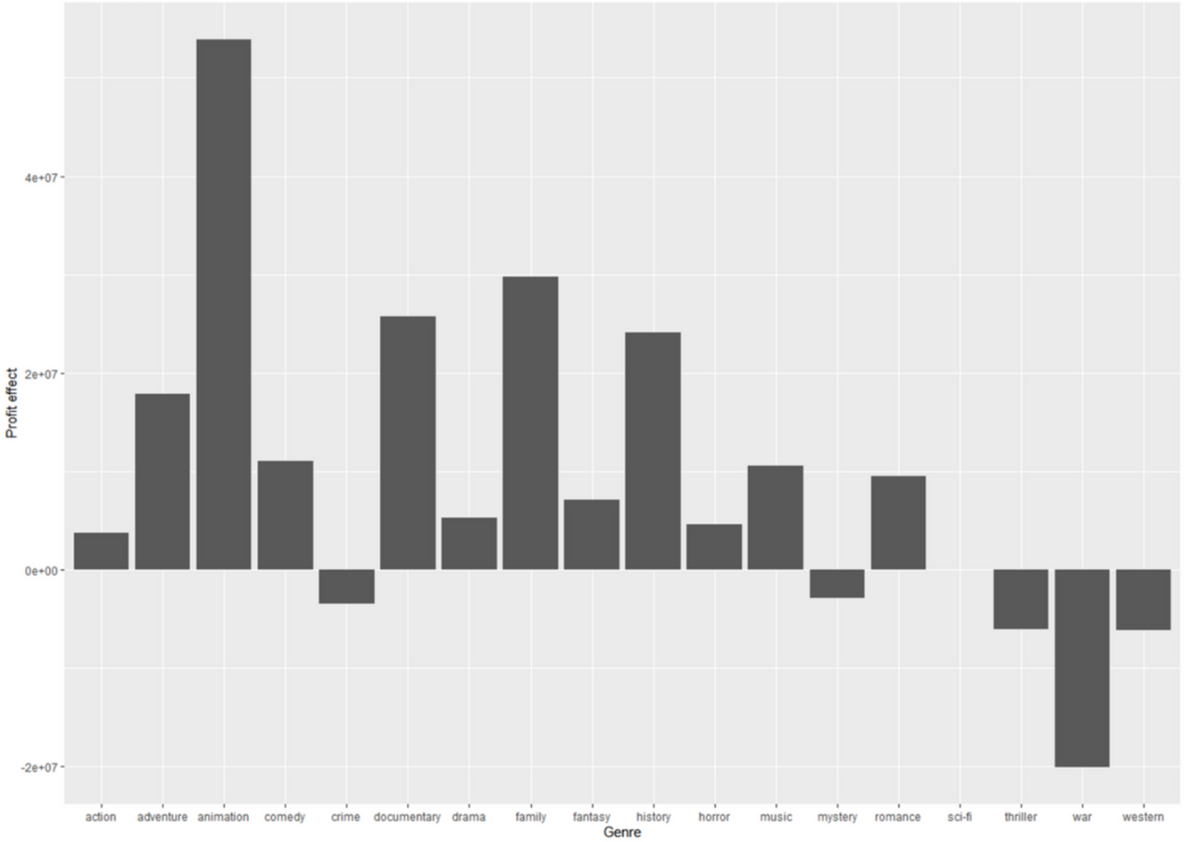
3. We chose to focus on **Warner and Universal Studios** because one had the lowest average profit and the other had high average profit

GROUP 1 - IB3K50

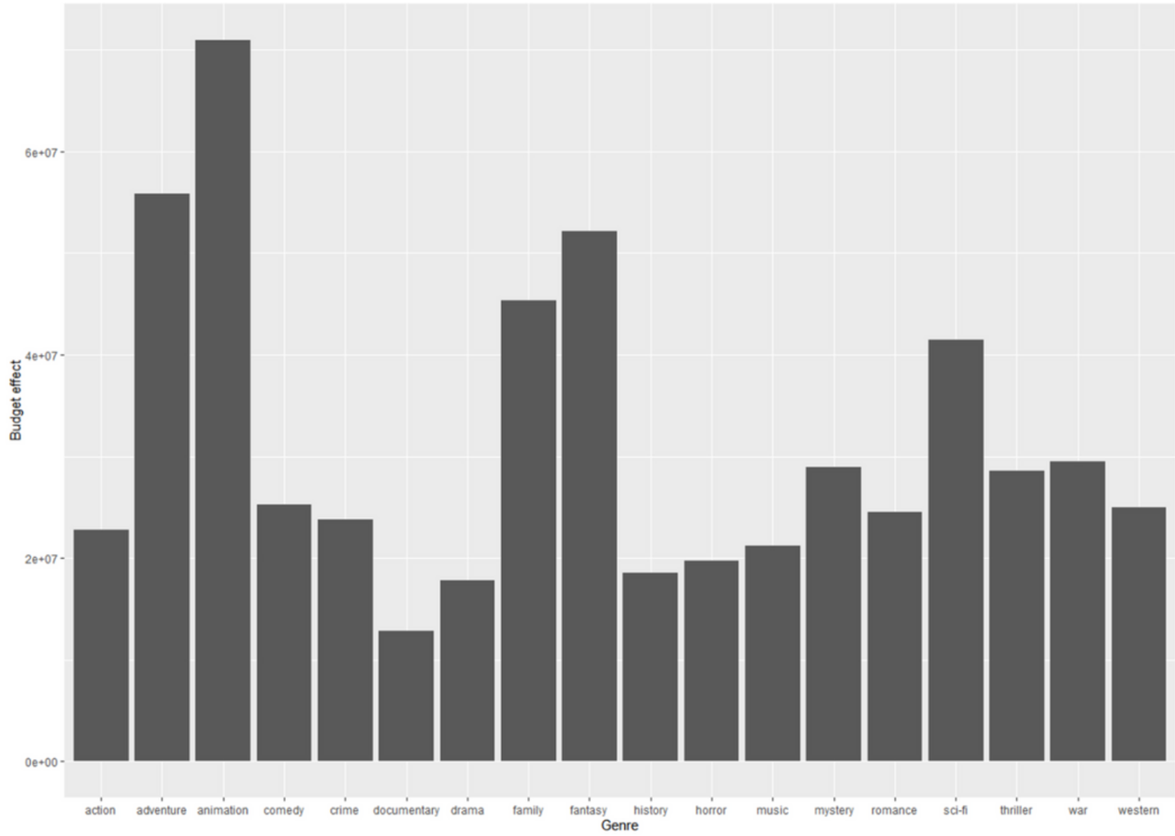


Analysis 1 : **Genre**

Profit Effect by Genre



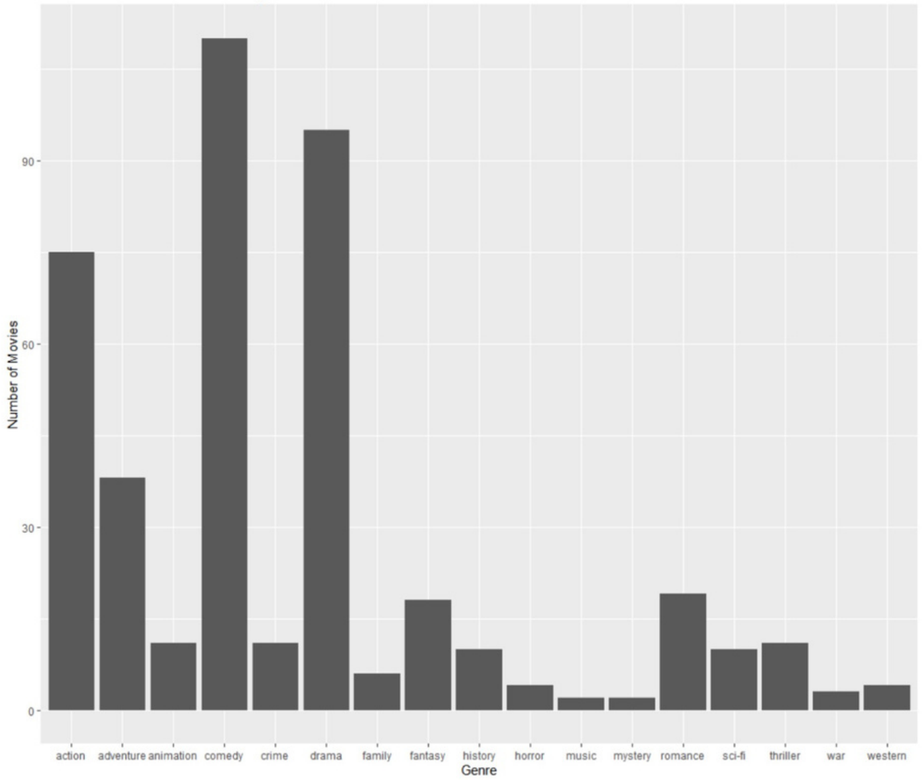
Budget Effect by Genre



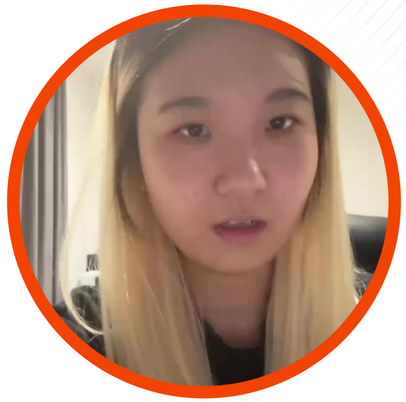
AI FOR BUSINESS

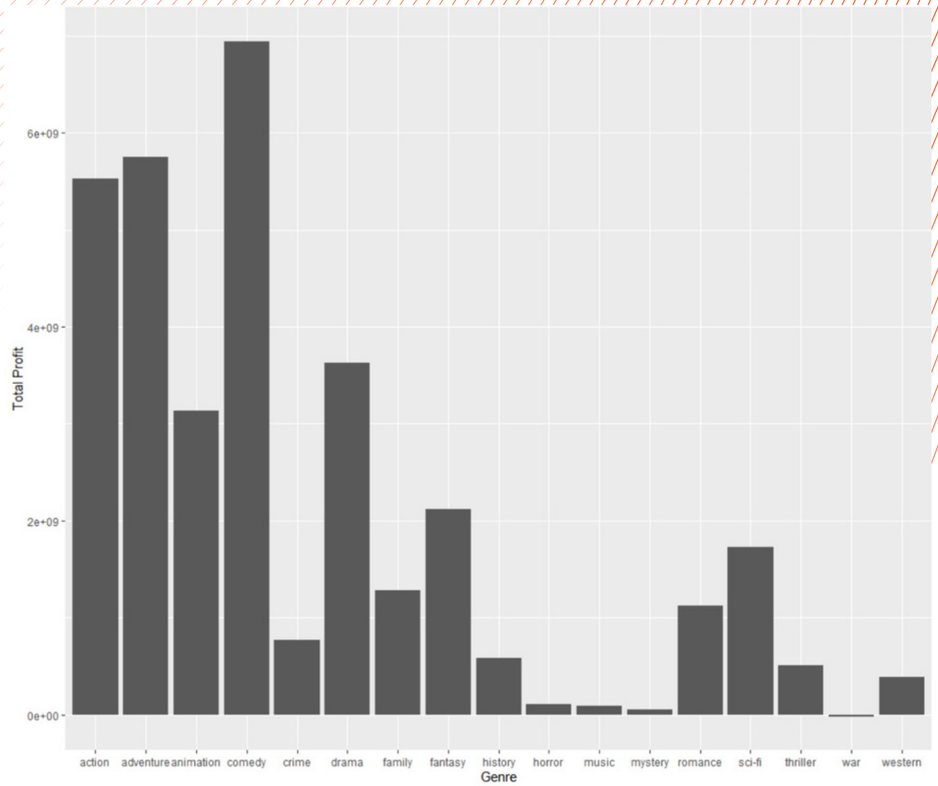
GROUP 1 - IB3K50

Universal: Specialised Genre



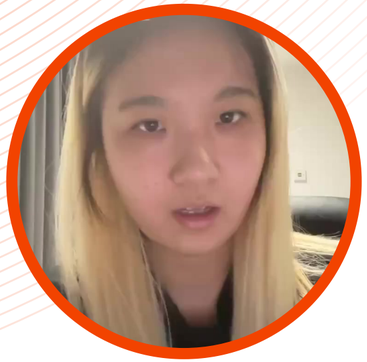
Mainly focusing on action, comedy and drama.
The profit effect is relatively low.



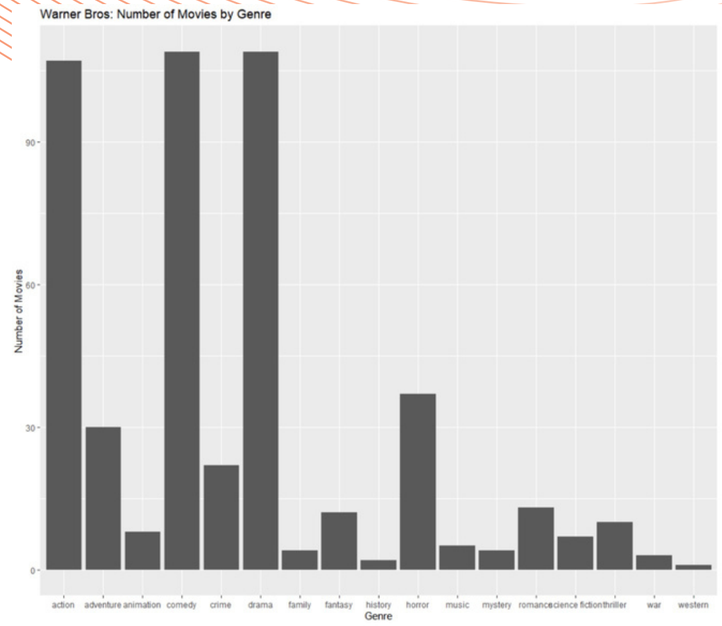


With numerous famous IPs, comedy has created the highest profit for Universal

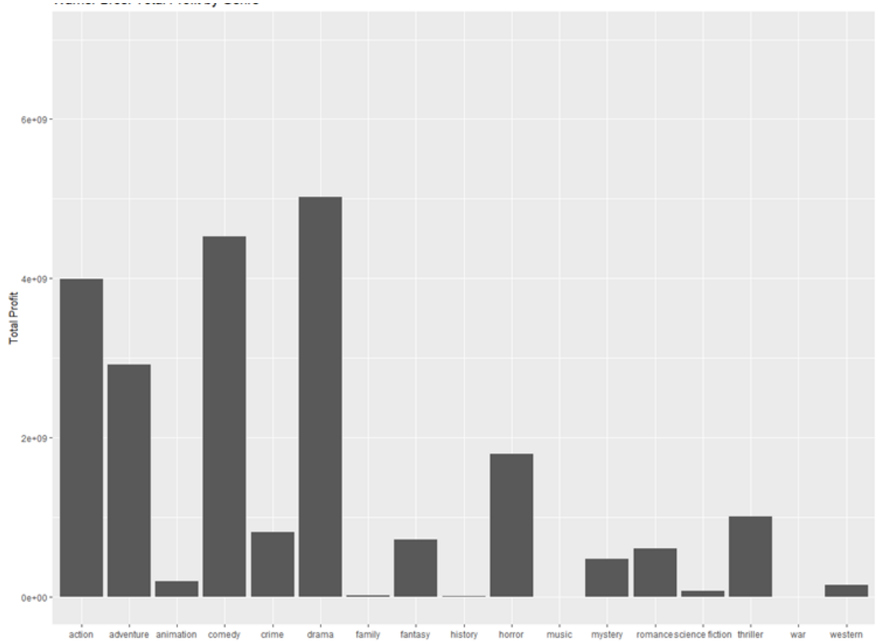
Warner Bros: Specialised Genre



Number of Movies by Genre



Total Profit by Genre

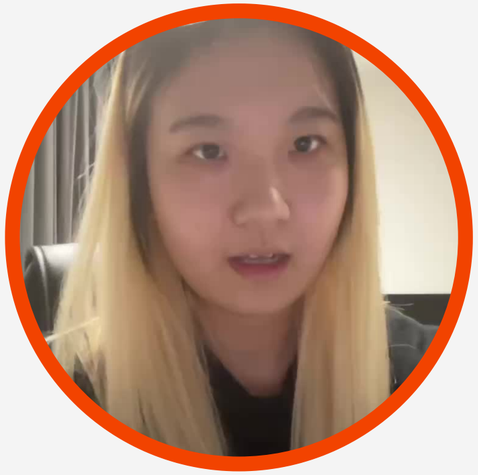


AI FOR BUSINESS

1. Focusing on **action, comedy and drama**
2. Compared to Universal, the drama took up a **higher percentage**
3. The target audience of **drama movies** is relatively narrower
4. The **design of content** would be more tricky to be attractive
5. **Less IPs** makes the story line poor

Summary of **Recommendations**

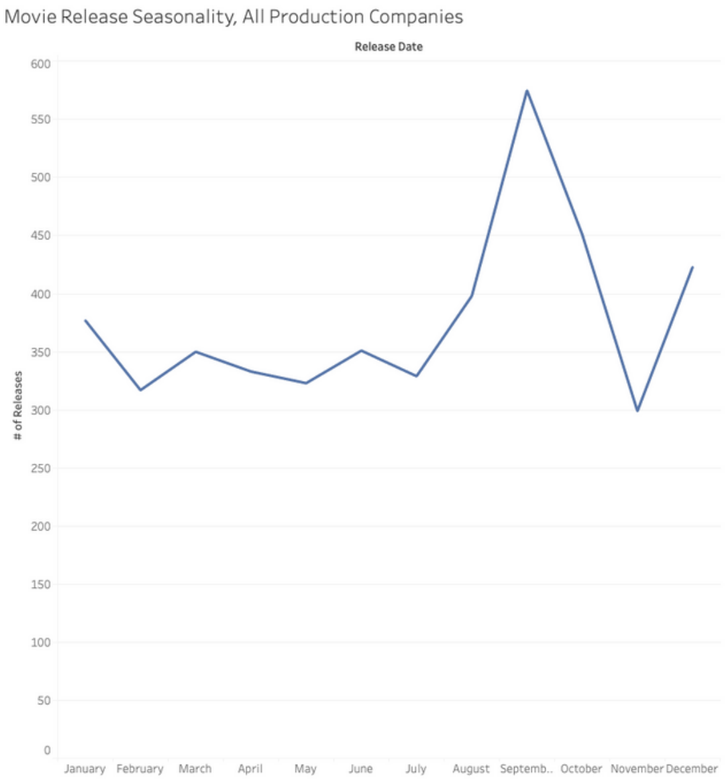
IP-related theme parks attract visitors with added value peripheral products and then transform them to **audiences.**



AI FOR BUSINESS

Movie Releases per Month

- **Spike of releases in September**, dip in Q3
- Constant level in Q1-Q2
- **January is 4th** month by # of releases



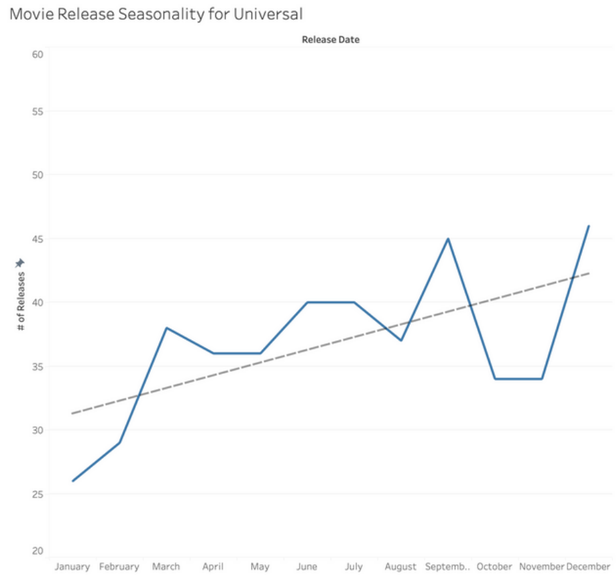
# of Movie Releases by Month, all Production Co's

1. # of releases by Universal and Warner **spike in December**, with a secondary, **lower spike in September**

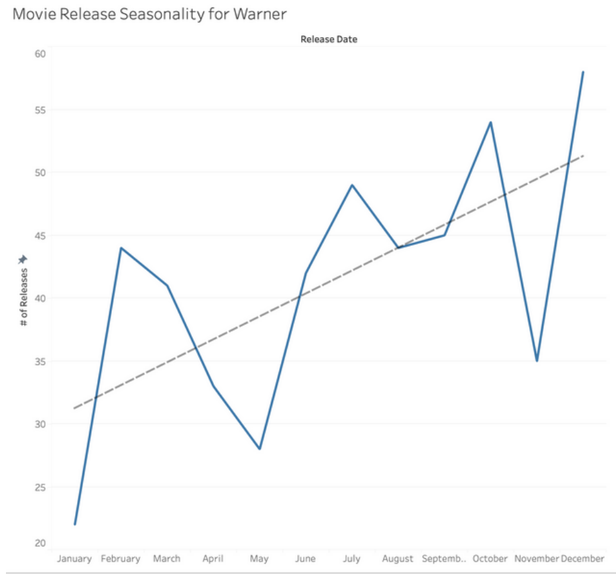
2. Both companies start with the **least # of releases in January** and **increase releases over the rest of the year**

3. Universal is able **maintain a higher average # of movies released per month**, whilst **Warner has higher variance**

of Movies by Month (Universal)

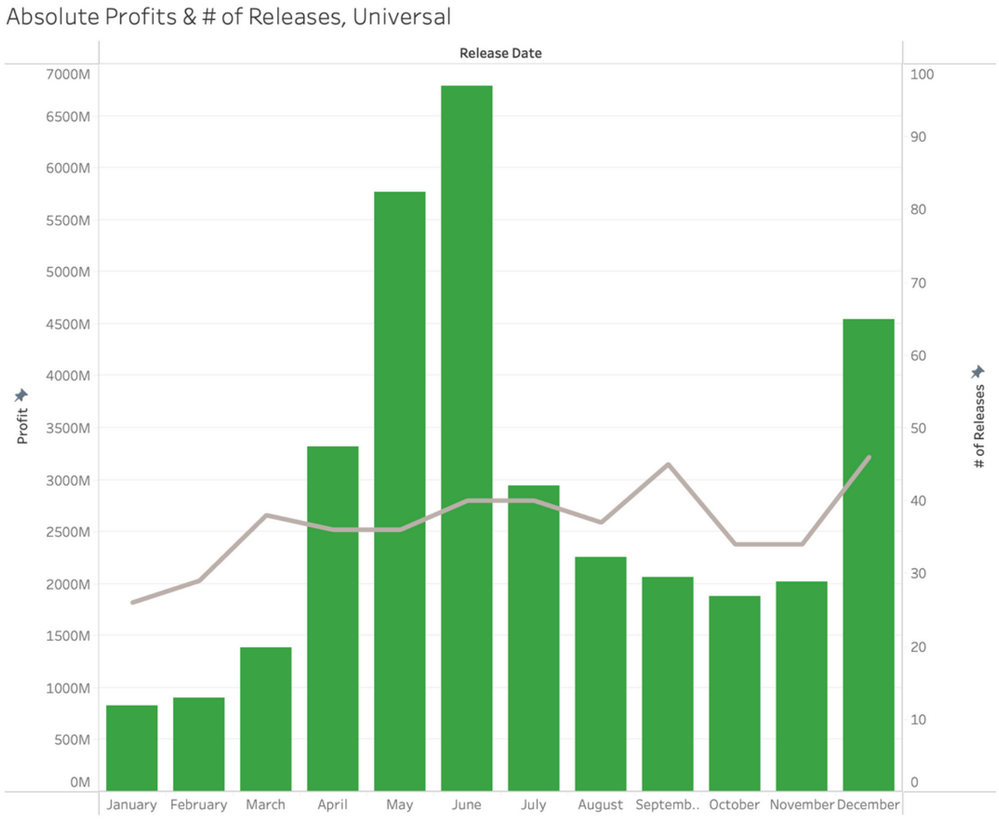


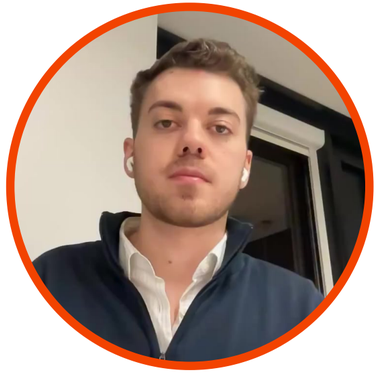
of Movies by Month (Warner Bros)



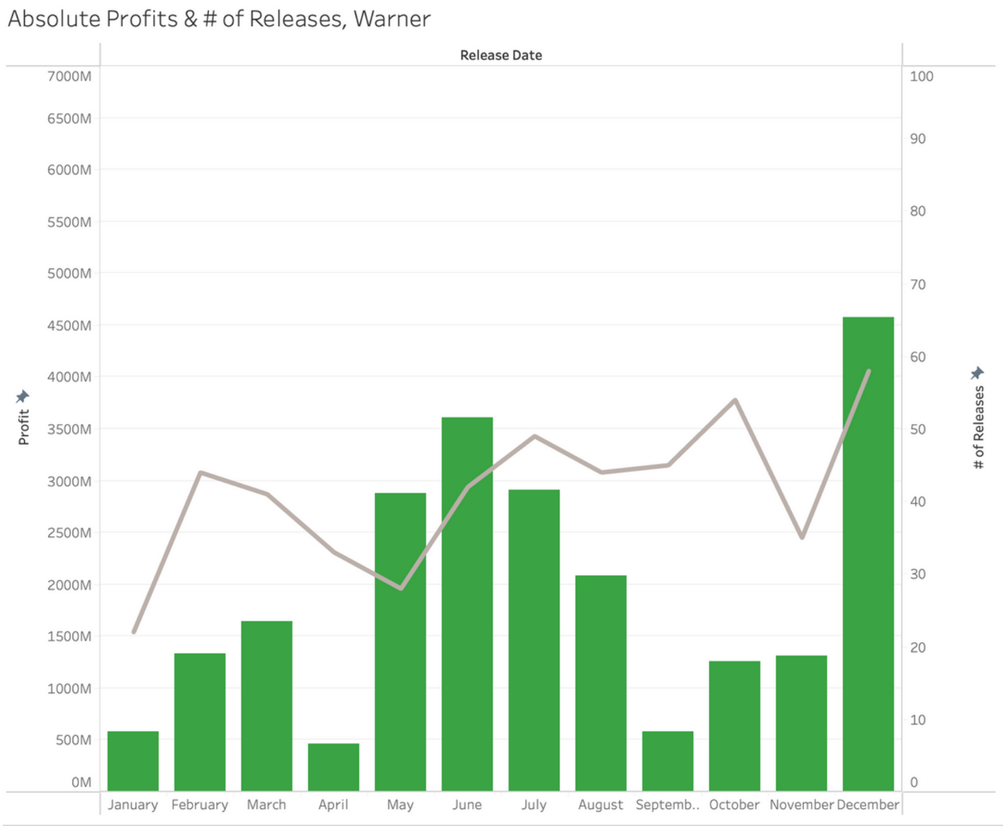
Trends from the chosen companies

Universal Studios: Absolute Profits & # of Releases





Warner Brothers: Absolute Profits & # of Releases



Implications:

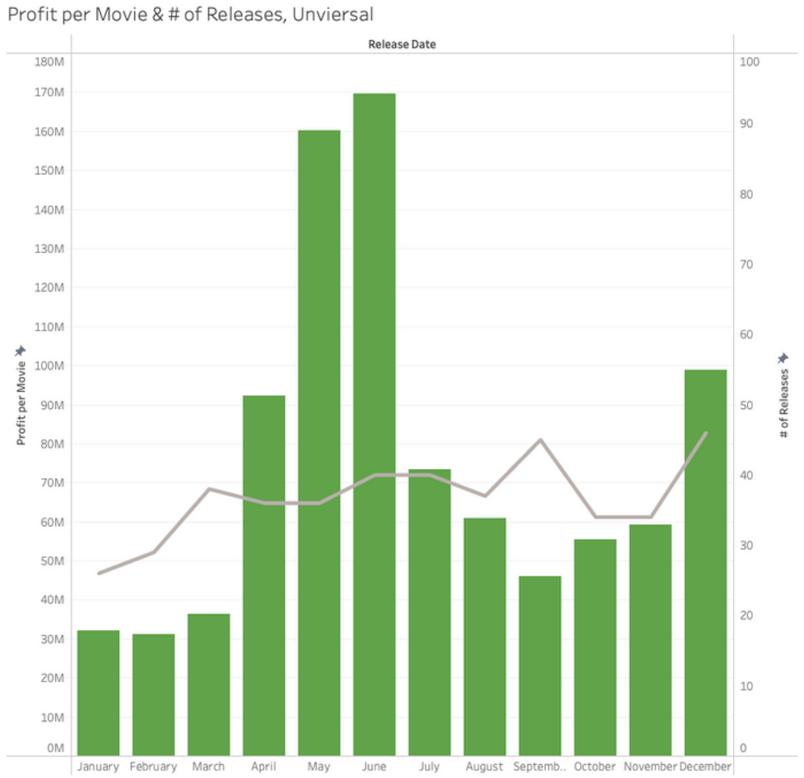
- High minimum profit per movie
- Highest profitability in Q2

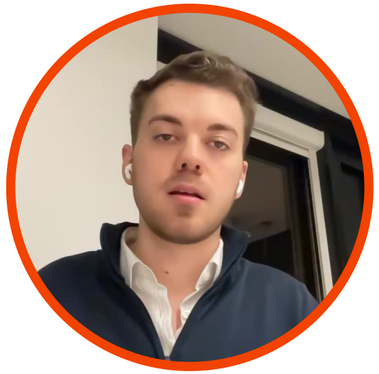
Implications:

- High variance in profits
- Q2 begins with worst month by profitability

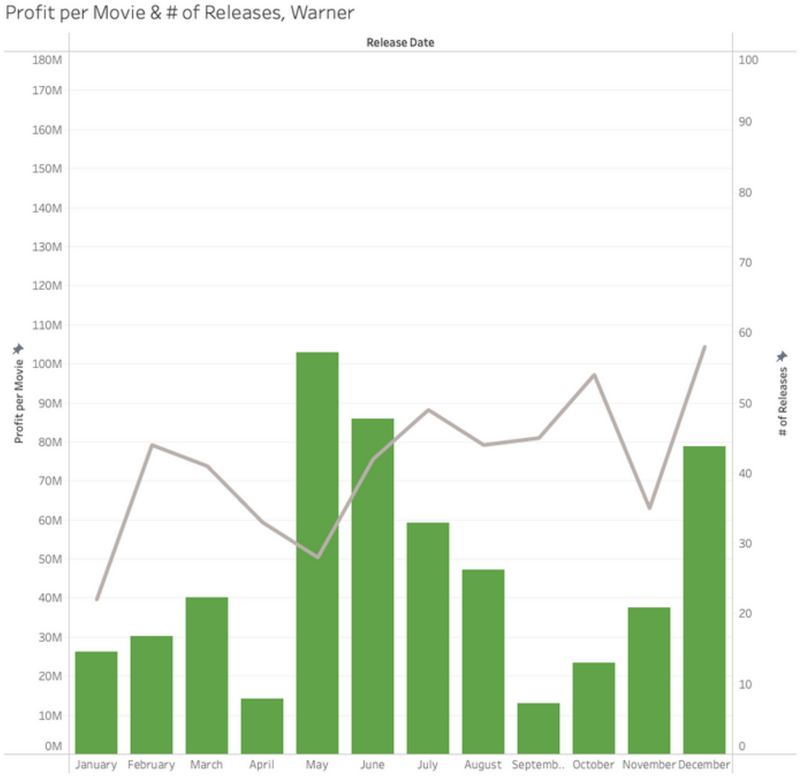
Trends from the chosen companies

Universal Studios: Profit per Movie





Warner Brothers: Profit per Movie



Implications:

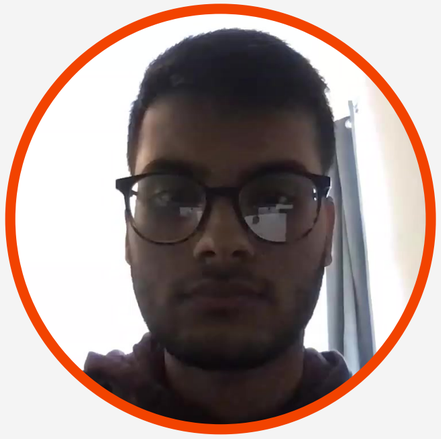
- Q2 remains strongest quarter by profitability
- Higher average profitability than Warner

Implications:

- Profitability lags Universal by 2 months
- Higher profit per movie is key to ensure long-term success

Popularity of Warner & Universal

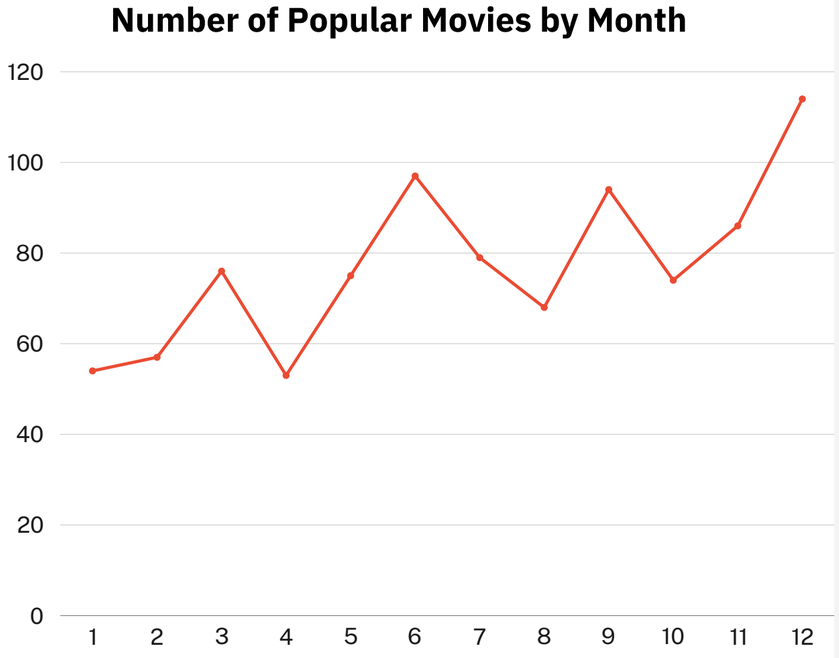
The months where movie popularity has peaked are **March**, **June**, **September**, and **December**.

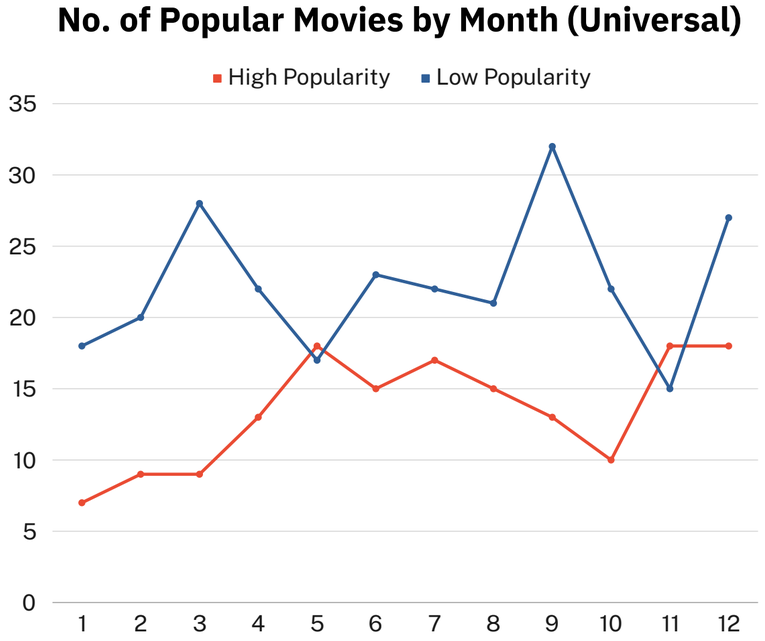


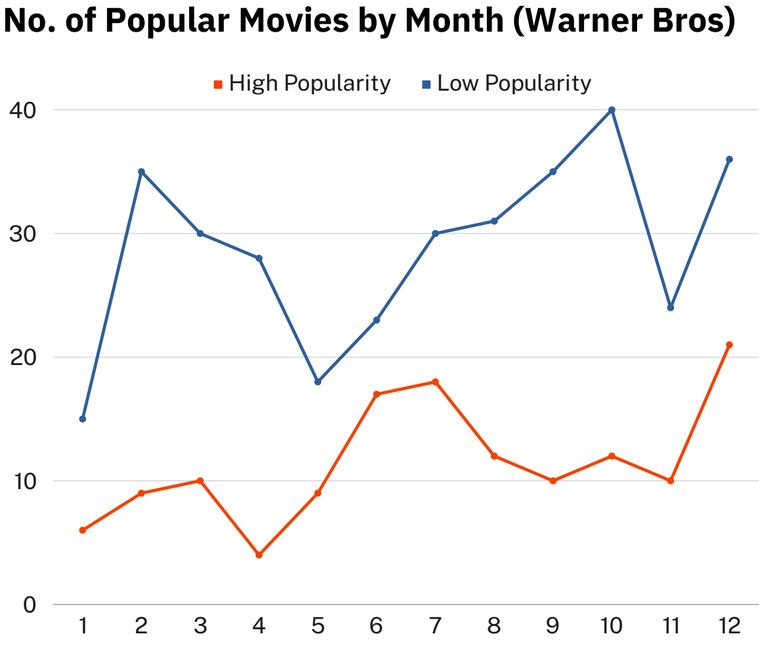
1. **Both** companies show similar trends in terms of releasing their popular and unpopular movies.

2. **Both** companies tend to increase the release of the overall number of movies in **December** regardless of popularity.

3. **Warner Bros** tends to release more **unpopular** movies

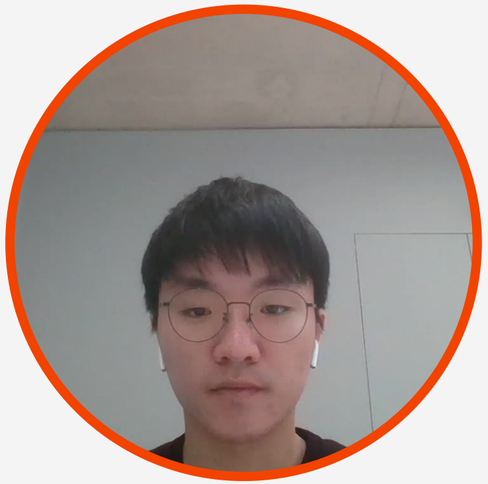






GROUP 1 - IB3K50

Popularity Prediction Model **Warner Bros**



Pop = Budget + Eng + Runtime + Vote_avg*Vote_cnt + Budget*Month

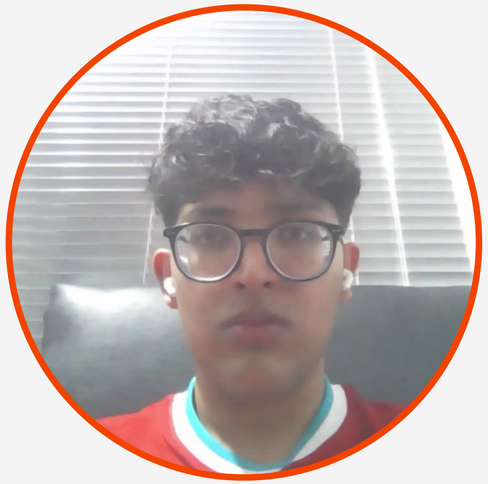
Predict the popularity of the movie before the actual release

Adj R squared: 0.761
RMSE: 9.185

Popularity	
Actual	Predict
61.20	69.62
53.91	76.13
44.93	40.59
50.31	59.49
24.86	28.27
50.77	37.04

GROUP 1 - IB3K50

Concluding Thoughts



AI FOR BUSINESS

Genre

- Universal focuses on comedy, which has a broader consumer range.
- Warner Bros specialises in drama, and does not have as many recognisable IPs.
- More profit for Universal

Seasonal Influence

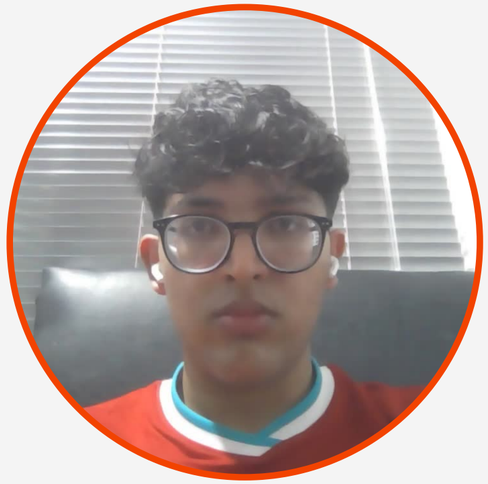
- December has the highest popularity due to being near the holiday period
- Warner Bros have a significant profit drop off during June

Popularity

- September a low profit month due to unpopular movies
- Factors such as financial planning are true cause of low profit in June
- Warner Bros have more unpopular movies in general

GROUP 1 - IB3K50

Recommendation



AI FOR BUSINESS

Increased Budget

Increased funds for marketing anf promotions would increase popularity

New approach to releases

- Release more movies in Decembet
- Decrease variance in number of releases per month
- Assess distribution methods

Diversify Genres

- Prevent average profitability from dropping
- Venture into areas like comedy to appeal to more consumers

Developing IPs

- Create franchises and continue sequels for existing successful films
- Attract loyal consumers, increase consistent engagement

GROUP 1 - IB3K50

Reflection: **Problems we faced**

Cleaning the Data

Combining Production Businesses

Inaccuracy?

Overlap -> Noise?

References

- The Numbers. Movie Production Companies. The Numbers. Available at: https://www.the-numbers.com/movies/production-companies/#production_companies_overview=l10:od3.
- Warner Bros. (2013). Pictures and RatPac-Dune Entertainment sign multi-year film financing agreement covering up to 75 titles, WarnerBros.com | Warner Bros. Pictures and RatPac-Dune Entertainment Sign Multi-year Film Financing Agreement Covering Up to 75 Titles | Press Releases. Warner Bros. Available at: https://www.warnerbros.com/news/press-releases/warner-bros-pictures-and-ratpac-dune-entertainment-sign-multi-year-film-financing.
- Reid, C. (2023). Warner's UK Studios Report Record Revenue. [online] Forbes. Available at: https://www.forbes.com/sites/carolinereid/2023/01/28/warners-uk-studios-report-record-revenue/?sh=4174195b1c33 [Accessed 4 Mar. 2023].
- Universal Pictures. (2013). Universal Pictures | About the Film Studio. [online] Available at: https://www.universalpictures.com/about [Accessed 8 Mar. 2023].
- Vorderer, P. (1993). Audience involvement and program loyalty. Poetics, 22(1-2), pp.89–98. doi:https://doi.org/10.1016/0304-422x(93)90022-9.
- Cho, H. and Eo, S.H., 2016. Outlier detection for mass spectrometric data. Statistical Analysis in Proteomics, pp.91-102.

Appendix 1: Regression Analysis and Scatter Plot

```
> library(readxl)
> movie <- read_excel("Desktop/Academics 22-23/movie.xlsx")
> View(movie)
> install.packages("fastDummies")
> library(fastDummies)
> data <- movie
> data <- dummy_cols(data,select_columns = "month")
> print(data)
> lmpopularitydatamonth = lm(popularity~budget
+ vote_average + vote_count + num_genre + runtime + month_2 + month_3
+ month_4 + month_5 +month_6 +month_7 +month_8+ month_9
+month_10 + month_11+month_12 , data = data)
> summary(lmpopularitydatamonth)
> install.packages("tidyverse")
> install.packages("ggplot2")
> library(ggplot2)
> ggplot(data = movie, mapping = aes(x = popularity, y = revenue)) + geom_point()
```

Appendix 2: Check for Entities

```r
##### extract companies
# clean production company variable
movie_company ← subset(movie, movie$production_companies≠"[]")
movie_company$production_companies2 ← purrr::map(movie_company$production_companies, jsonlite::fromJSON)

# extract production company to separate dataset
company ← data.frame(name = "universal pictures")
for (i in seq(1:nrow(movie_company))){
  for (j in seq(1:nrow(movie_company$production_companies2[[i]]))){
    company ← company %>% add_row(name = movie_company$production_companies2[[i]][[1]][[j]])
  }
}
company ← distinct(company)

# check for entities
entities ← subset(company, grepl("century fox", company$name))
```

Appendix 3: Prediction Model R Code

```r
########## predict popularity ##########
##### step 1
set.seed(12123)
trainingRowIndex ← sample(1:nrow(warner), 0.8*nrow(warner))

trainingData ← warner[trainingRowIndex,]
dim(trainingData)

testData ← warner[-trainingRowIndex,]
dim(testData)

##### step 2
popularity_model ← lm(popularity ~ budget + english + runtime + vote_average*vote_count + budget*month, data = trainingData)
summary(popularity_model)

movie1 ← testData[4,]
print(movie1$popularity)

movie1_pred ← predict(popularity_model, newdata = movie1)
movie1_pred

popularity_predict1 ← predict(popularity_model, newdata = testData)

actual_preds1 ← data.frame(cbind(actual_popularity = testData$popularity, predicted_popularity = popularity_predict1))
head(actual_preds1)

rmse(actual_preds1$actual_popularity, actual_preds1$predicted_popularity)
# 9.185276
```

Appendix 4: Regression in Prediction Model

```
Call:
lm(formula = popularity ~ budget + english + runtime + vote_average *
    vote_count + budget * month, data = trainingData)

Residuals:
    Min      1Q  Median      3Q     Max
-40.166  -4.336  -0.471   4.008  29.400

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.247e+01  5.236e+00  -2.382   0.0177 *
budget                   1.288e-09  2.801e-08   0.046   0.9634
english                  3.592e+00  3.713e+00   0.967   0.3339
runtime                  6.985e-03  1.826e-02   0.383   0.7023
vote_average             2.992e+00  5.851e-01   5.113 5.06e-07 ***
vote_count               4.300e-02  4.771e-03   9.013  < 2e-16 ***
month                   -1.356e-01  1.699e-01  -0.798   0.4256
vote_average:vote_count -3.929e-03  6.764e-04  -5.809 1.34e-08 ***
budget:month             4.277e-09  3.271e-09   1.307   0.1919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.351 on 377 degrees of freedom
Multiple R-squared:  0.7659,    Adjusted R-squared:  0.7609
F-statistic: 154.2 on 8 and 377 DF,  p-value: < 2.2e-16
```