Hit Movie Project

Group 8

- John McDowell - “Square”
- Jeremy Ocain - “Circle”
- Latifou Amoussa - “Triangle”
- Jose Mendez - “X”

Why Hit Movies?

- We all enjoy movies, but what makes them hits?
- What is our definition of a “hit”?
- This analysis attempts to dissect what is a critical hit based on Metascore data and a Monetary hit based on Net Profit.

Data Sources and Tools Utilized

1. CSV file from Kaggle - IMDb movies.csv
2. Pandas / Python / Sqlalchemy / Scikit-learn
3. SQLite to clean and integrate data
4. Tableau for visualizations and final presentation

Purpose

- Is a movie a hit based on simply Metascore, Gross Income or a mixture of these outputs?
- Is there a seasonality effect in place?
- Does higher budget increase hit probability?
- Has there been a significant change in profitability for movies over the decades analyzed?
- Is there any correlation by genres?

Definitions / Terminology Used

Critically Acclaimed Movies (Based off of Metascore)

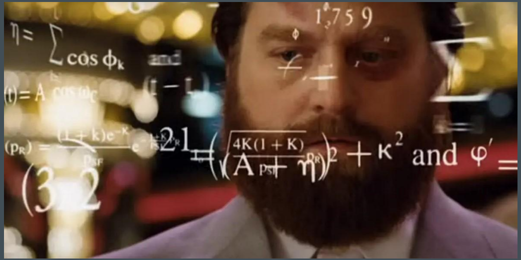
Our group considers a movie to be critically acclaimed (meta_hit) if the movie achieves a metascore of 75 or greater.

Blockbuster movies

Our group considers a blockbuster movie by two stipulations:

- A movie with a budget less than 7 million and having a gross profit greater than or equal to 500 percent.
- A movie with a budget over 7 million and having a gross profit greater than or equal to 250 percent.

The movie data utilized in this project span from the 1980’s to the beginning of 2020



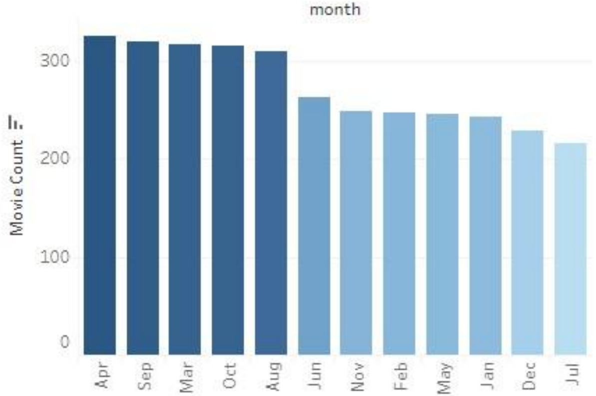
Total Movie Count and Net Profits

What makes a Hit Movie?

‹ 1 2 3 4 5 6 ›

Year (date) (group)
(All)

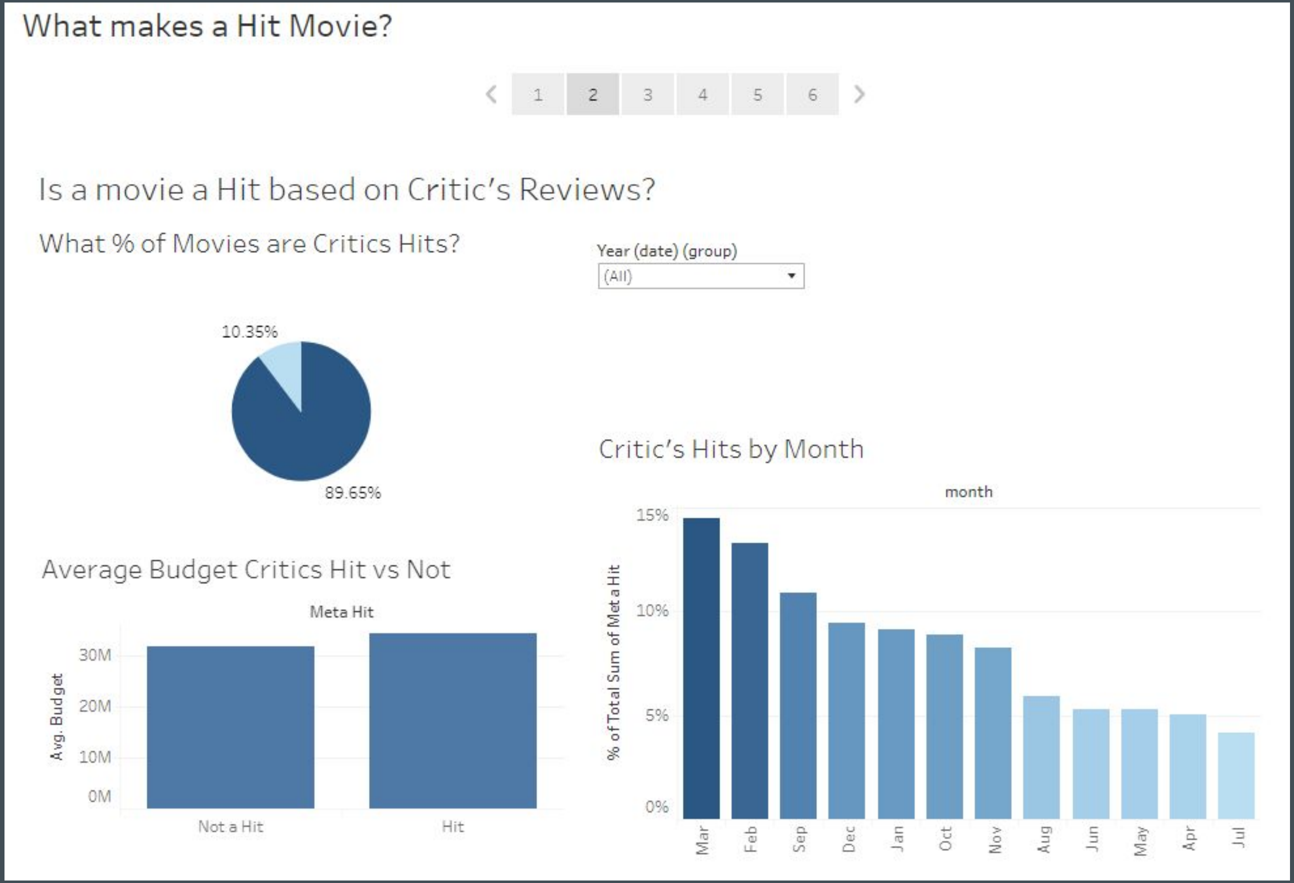
Total movies by month



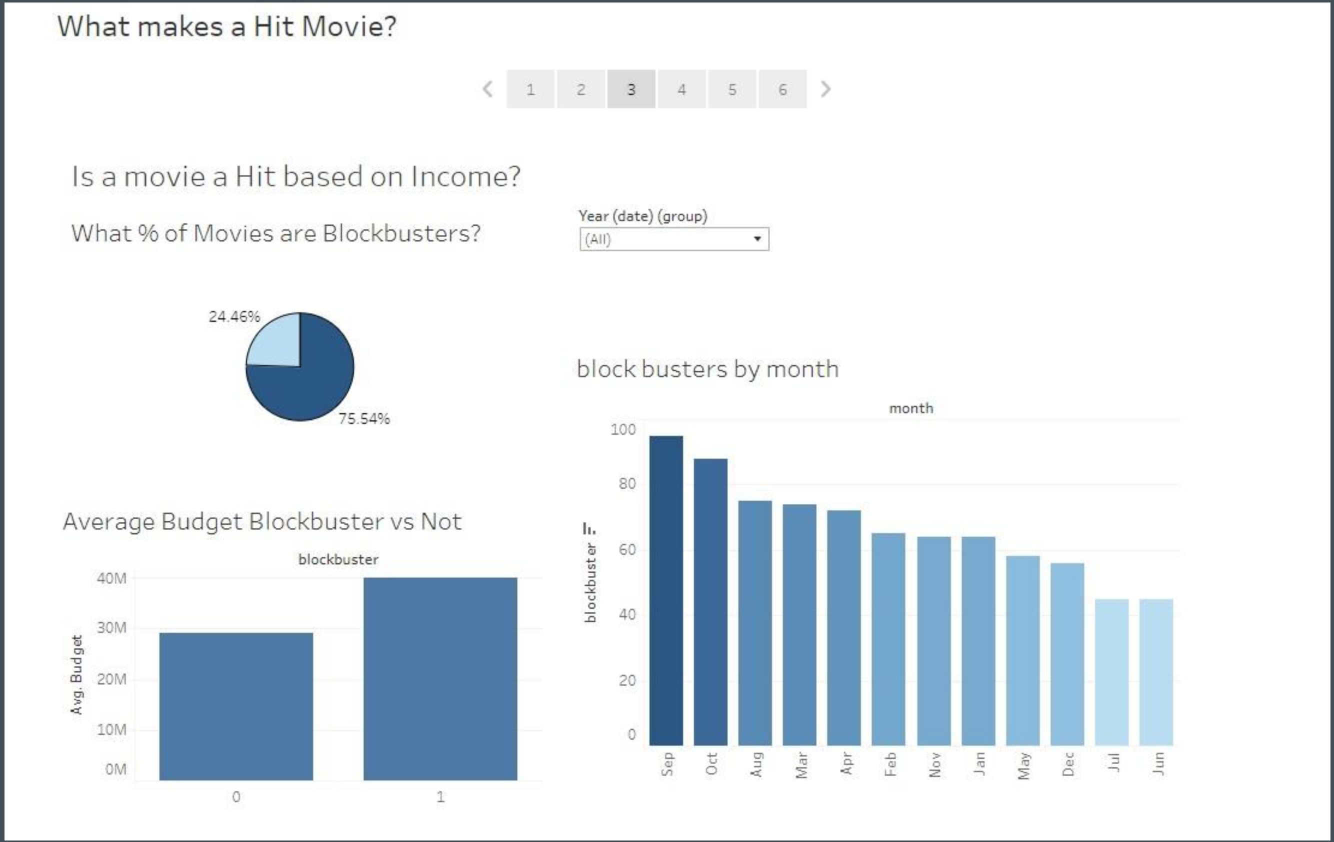
Net Profits by Month

month	
Dec	24,444,711,152
Sep	24,048,720,563
Oct	22,937,406,371
Apr	19,982,549,257
Nov	17,548,030,944
Aug	16,923,874,503
Mar	15,771,014,830
Jan	15,260,219,345
May	15,185,191,105
Feb	13,914,909,765
Jun	13,790,053,036
Jul	8,158,954,406

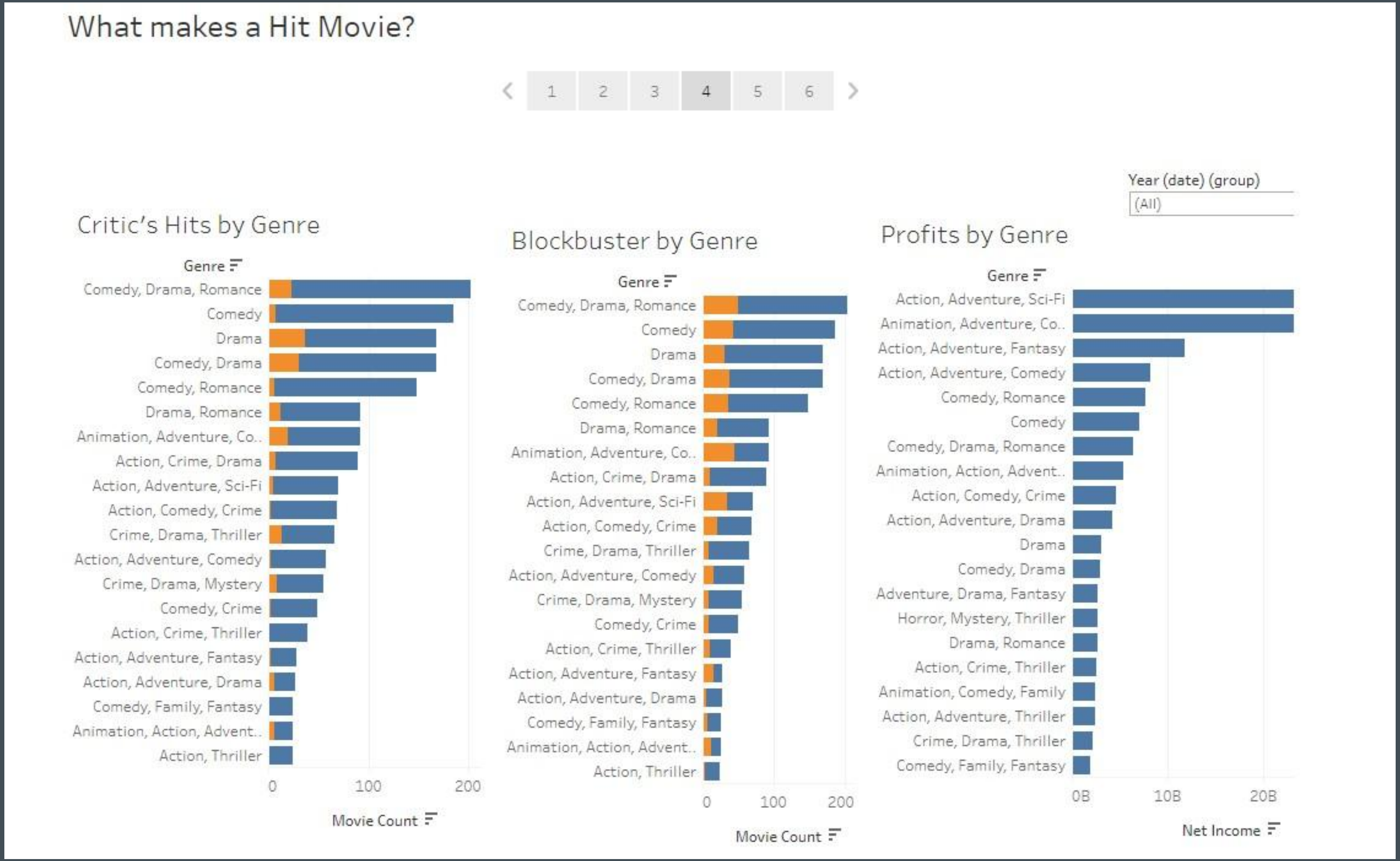
Critical Hits - Budget and Month Data



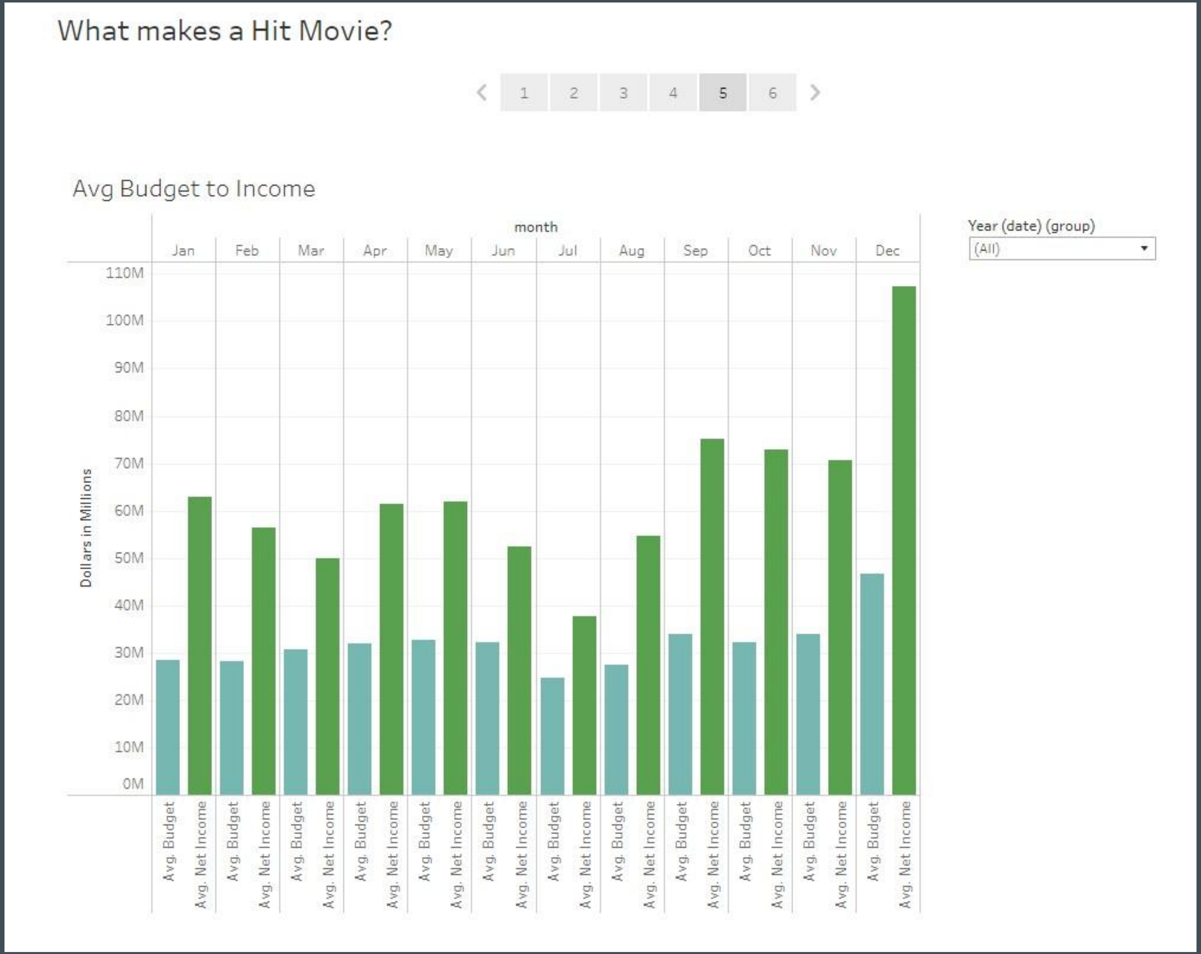
Block-Buster Analysis



Hit Movie by Genre Analysis



Budget Analysis



Initial Observations from Tableau

Key Observations from decades analysis:

- Most movies come out in April, September, March and October. Profitability reflects the same pattern.
- The most critically acclaimed movies come out in March and February and are a combination of Drama, Comedy and Romance genres.
- The biggest money makers come out in September and October and are a combination of Action, Adventure, Sci-Fi, and Comedy genres.
- The one month that is significantly different from the rest is December. The fewest movies are released, but make the most profits.

Machine Learning - Part 1

Machine Learning:

- Based off our group's desired outcomes being binary, whether a movie will be a hit or not, we decided that using logistic regression models would be best.
- From the visualizations we determined that the strongest correlations are between the movie's budget, total gross, net income, Metascore (based off critic reviews), as well as both month and genre.
- After preprocessing our data and creating classification algorithms for the desired outcomes we created 6 models to accurately predict whether a movie will be a hit for future producers and directors.

Machine Learning - Part 2

Critically Acclaimed Movies:

	Predicted Critical Hit	Predicted Non-Critical Hit
Actual Critical Hit	719	28
Actual Non-Critical Hit	76	10

Predictabily of a movie being critically acclaimed

```
report = classification_report(y_test, y_pred)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.90      0.96      0.93       747
           1       0.26      0.12      0.16        86

    accuracy                           0.88       833
   macro avg       0.58      0.54      0.55       833
weighted avg       0.84      0.88      0.85       833
```

	Predicted Critical Hit	Predicted Critical Hit
Actual Critical Hit	731	3
Actual Non-Critical Hit	83	2

Predictabily of a movie being a critically acclaimed movie based on genre

```
report = classification_report(y_test, y_pred)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.90      1.00      0.94       734
           1       0.40      0.02      0.04        85

    accuracy                           0.89       819
   macro avg       0.65      0.51      0.49       819
weighted avg       0.85      0.89      0.85       819
```

	Predicted Critical Hit	Predicted Critical Hit
Actual Critical Hit	731	3
Actual Non-Critical Hit	83	2

Predictability of a movie being critically acclaimed based off release month

```
report = classification_report(y_test, y_pred)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.90      1.00      0.94       734
           1       0.40      0.02      0.04        85

    accuracy                           0.89       819
   macro avg       0.65      0.51      0.49       819
weighted avg       0.85      0.89      0.85       819
```

Machine Learning - Part 3

Blockbuster Movies:

	Predicted Critical Hit	Predicted Critical Hit
Actual Critical Hit	614	5
Actual Non-Critical Hit	20	180

Predictability of movie being a blockbuster based on its release month

```
report = classification_report(y_test, y_pred)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       619
           1       0.97      0.90      0.94       200

    accuracy                           0.97       819
   macro avg       0.97      0.95      0.96       819
weighted avg       0.97      0.97      0.97       819
```


	Predicted Critical Hit	Predicted Critical Hit
Actual Critical Hit	614	5
Actual Non-Critical Hit	20	180

Predictability of movie being a blockbuster based on its release month

```
report = classification_report(y_test, y_pred)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       619
           1       0.97      0.90      0.94       200

    accuracy                           0.97       819
   macro avg       0.97      0.95      0.96       819
weighted avg       0.97      0.97      0.97       819
```


	Predicted Blockbuster Hit	Predicted Non-Blockbuster
Actual Blockbuster Hit	614	5
Actual Non-Blockbuster	20	180

Predictability of a movie being a blockbuster hit based on its genre

```
report = classification_report(y_test, y_pred)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       619
           1       0.97      0.90      0.94       200

    accuracy                           0.97       819
   macro avg       0.97      0.95      0.96       819
weighted avg       0.97      0.97      0.97       819
```

Conclusions

- Net Profits by Month - Although December shows to be the most profitable month historically spend, the trend for the last 20 years shows April to be the most profitable month.
- The most critical hits come from February and March months, near Academy Awards season.
- There are 33% less movies considered critical hits from the 1980's and 1990's to the 2000's and 2010's.
- The budget for blockbusters has ballooned by an average of 3-fold over the last 20 years (2000s / 2010s) when comparing movies from the (1980's / 1990's).
- Accuracy Score for MetaHit Logistic Regression Model is 0.8955582232893158
- Accuracy Score for Blockbuster Logistic Regression Model is 0.9682539682539683

Any Questions?
Thank you for your attention!

Data Set Reference:
Leone, S. (2020). *IMDb movies extensive dataset* (Version 2) [Data set].
https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv