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.

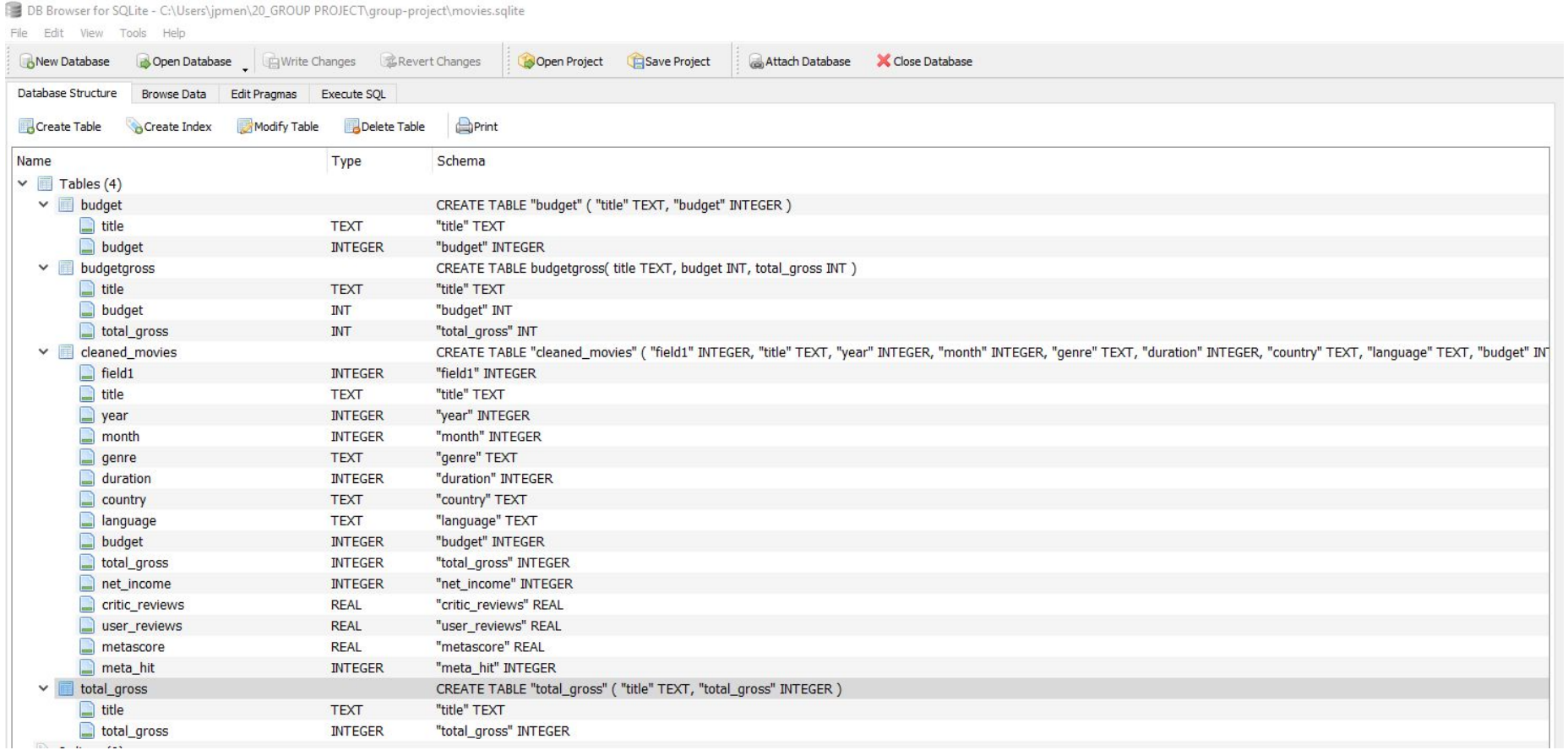
Data Sources

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?

Building the Database - Data Exploration



Name	Type	Schema
Tables (4)		
budget		CREATE TABLE "budget" ("title" TEXT, "budget" INTEGER)
title	TEXT	"title" TEXT
budget	INTEGER	"budget" INTEGER
budgetgross		CREATE TABLE budgetgross(title TEXT, budget INT, total_gross INT)
title	TEXT	"title" TEXT
budget	INT	"budget" INT
total_gross	INT	"total_gross" INT
cleaned_movies		CREATE TABLE "cleaned_movies" ("field1" INTEGER, "title" TEXT, "year" INTEGER, "month" INTEGER, "genre" TEXT, "duration" INTEGER, "country" TEXT, "language" TEXT, "budget" INT
field1	INTEGER	"field1" INTEGER
title	TEXT	"title" TEXT
year	INTEGER	"year" INTEGER
month	INTEGER	"month" INTEGER
genre	TEXT	"genre" TEXT
duration	INTEGER	"duration" INTEGER
country	TEXT	"country" TEXT
language	TEXT	"language" TEXT
budget	INTEGER	"budget" INTEGER
total_gross	INTEGER	"total_gross" INTEGER
net_income	INTEGER	"net_income" INTEGER
critic_reviews	REAL	"critic_reviews" REAL
user_reviews	REAL	"user_reviews" REAL
metascore	REAL	"metascore" REAL
meta_hit	INTEGER	"meta_hit" INTEGER
total_gross		CREATE TABLE "total_gross" ("title" TEXT, "total_gross" INTEGER)
title	TEXT	"title" TEXT
total_gross	INTEGER	"total_gross" INTEGER

Hit v. Non-Hit Blockbuster - Data Exploration

In [17]:
```
# Calculate the balanced accuracy score
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, y_pred))
```

```
0.9694749694749695
```

In [18]:
```
# Display the confusion matrix
from sklearn.metrics import confusion_matrix, classification_report
matrix = confusion_matrix(y_test, y_pred)

# create a dataframe from the confusion matrix
matrix_df = pd.DataFrame(matrix, index=['Actual Blockbuster Hit', 'Actual Non-Blockbuster'], columns=['Predicted Blockbuster
matrix_df
```

Out[18]:

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

In [22]:
```
report = classification_report(y_test, y_pred)
print(report)
```

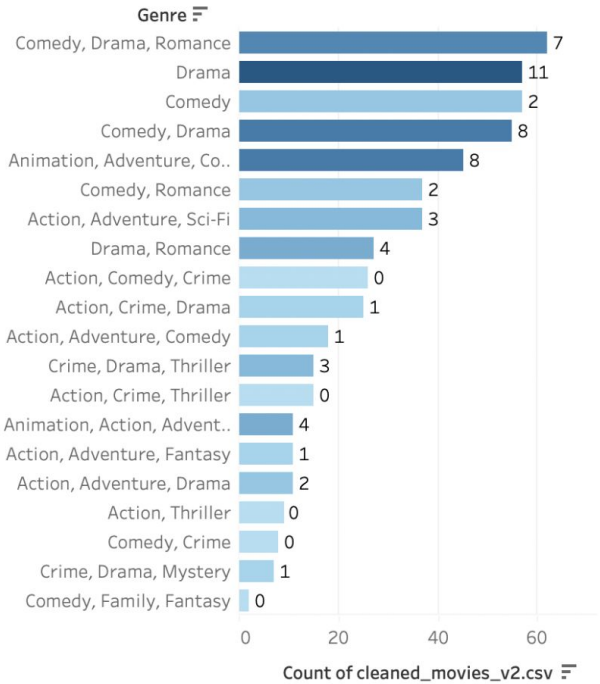
```
              precision    recall  f1-score   support

           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
```

Critical Hits and Profits by Genre (2010's) - Description of Analysis Phase

Critic's Hits by Genre



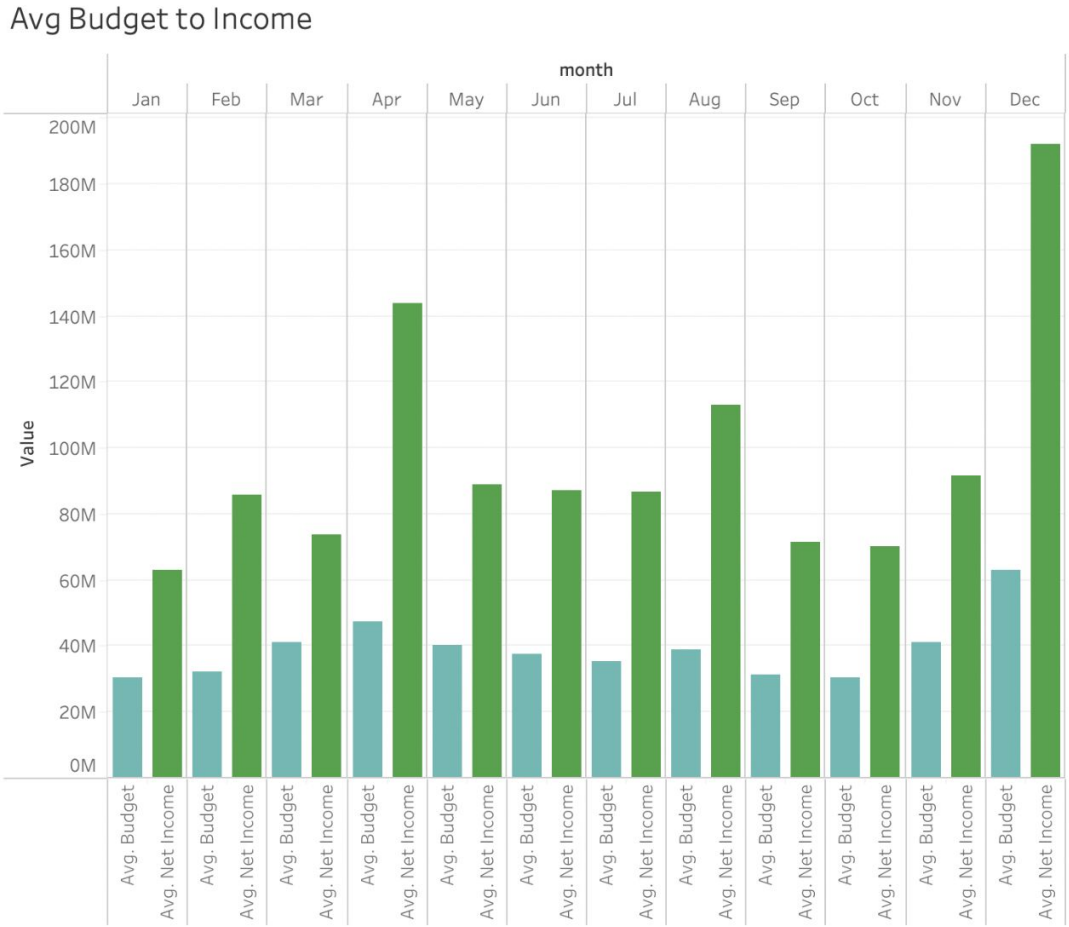
Profits by Genre

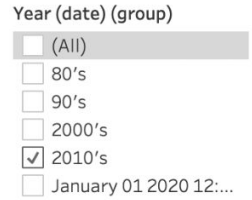
Genre	
Action, Adventure, Sci-Fi	21,881,144,623
Animation, Adventure, Co..	17,646,227,091
Action, Adventure, Comedy	5,822,274,311
Action, Adventure, Fantasy	3,976,220,801
Animation, Action, Adven..	3,760,409,911
Action, Adventure, Drama	3,300,855,113
Comedy	3,204,985,334
Horror, Mystery, Thriller	2,599,905,851
Comedy, Romance	2,114,254,271
Adventure, Drama, Fantasy	1,961,751,769
Action, Comedy, Crime	1,765,966,269
Comedy, Drama, Romance	1,586,929,762
Comedy, Drama	1,284,445,617
Drama, Romance	917,735,243
Animation, Comedy, Family	857,899,784
Action, Crime, Thriller	565,479,593
Drama	556,997,819
Crime, Drama, Thriller	397,204,413
Comedy, Family, Fantasy	222,214,513
Action, Adventure, Thriller	70,499,399
Grand Total	74,493,401,487

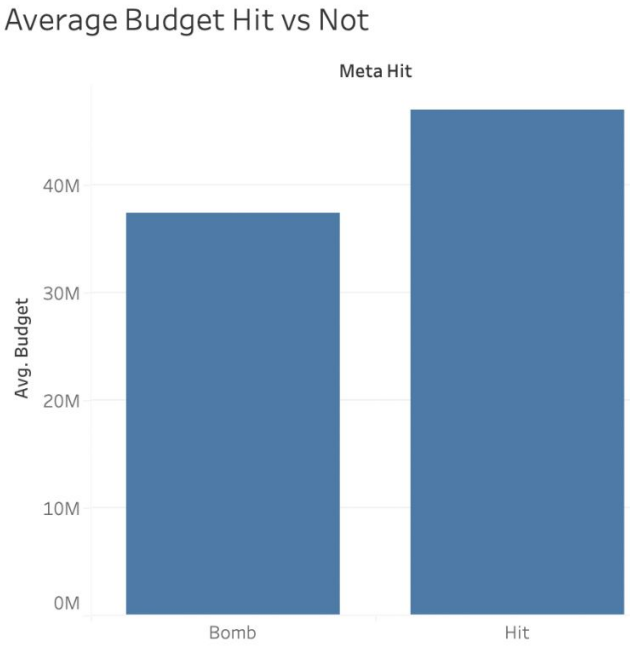
Year (date) (group)

- [ ] (All)
- [ ] 80's
- [ ] 90's
- [ ] 2000's
- [x] 2010's
- [ ] January 01 2020 12:...

Budget to Income (2010's) - Description of Analysis Phase

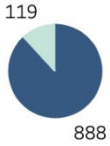






Movie Releases / Hits / Net Profits by Month (2010's) - Description of Analysis Phase

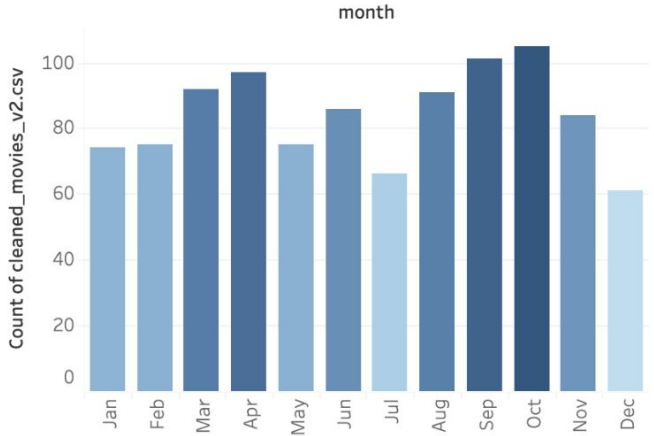
How many hit movies?



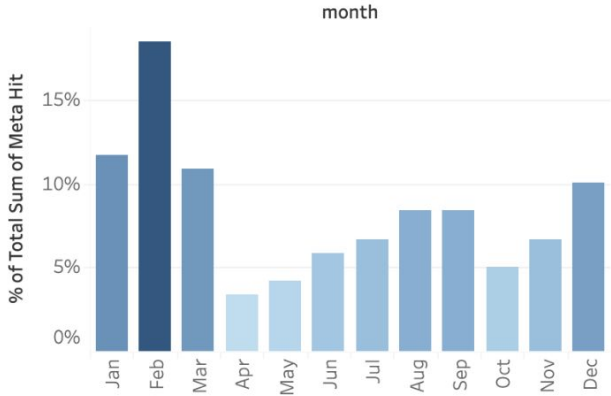
Year (date) (group)

2010's

movies by month



Critic's Hits by Month



Net Profits by Month

month	
Apr	13,935,332,428
Dec	11,704,176,505
Aug	10,296,805,534
Nov	7,688,921,810
Jun	7,506,965,799
Oct	7,366,945,939
Mar	6,790,540,029
Sep	7,212,656,740
May	6,655,061,273
Feb	6,440,190,592
Jul	5,719,428,740
Jan	4,663,269,365

Budget v. Net Profit by Month (2010's) - Description of Analysis Phase