Scope

To use databases and supervised machine learning to predict if a movie will be a hit or not.

Chose this topic because we all enjoy watching movies and prefer movies that are 'hits'.

Utilized the following items for this “hit movie project".

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 google slide for final presentation

Communication Protocols

- Created a group in Slack with all 4 members to discuss the project.
- Use Zoom when needed to discuss the project, share screen, debug code, and assist teammates.

Questions we hope to answer with this analysis

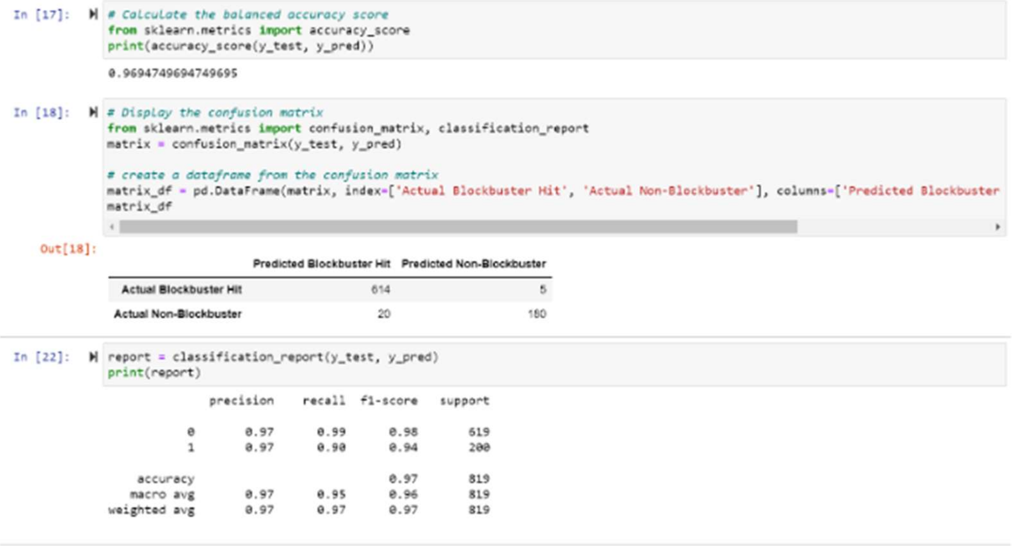
1. Is a movie a hit based on its Metascore?
2. Is a movie a hit based on how much money on total gross income?
3. Will a movie be a hit depending on season of the year it was released?

Steps takes to clean and to Analyze the Data

1. Dropping unnecessary columns.
2. Replacing None values with NaN
3. Filling null duration values with the mean duration.
4. Dropping NaN values.
5. Keeping movies only from 1980 and after.
6. Keeping movies only made in USA.
7. Changing columns with monetary values to numeric and dropping the dollar signs.
8. Extracting month from the date_published column and creating a new month column.
9. Dropping the date_published column.
10. Creating a net income column.
11. Rearranging columns in a more logical order.
12. Converting columns with object data types to float.
13. Creating a new meta_hit column for movies that have a Metascore of 75 or greater.
14. Convert new meta_hit column to numeric.
15. Saving new cleaned, preprocessed data to a new csv.

We chose to use Metascore for our target because the Metascore is based on critics' reviews and tells if a movie is critically acclaimed. Therefore, the data was split into training and testing sets based on whether it was a 'meta_hit.' We also chose a logistic regression model because we want to predict the likelihood of a movie being critically acclaimed in the future. Since we have a binary dependent variable of meta_hit, this was the best model for our project. The model had a very high accuracy score of 90% and had >0.90

precision, recall, and f1 scores on whether a movie was not a critical hit. However, there were very low precision, recall, and f1 scores on whether a film would be a critical hit.



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

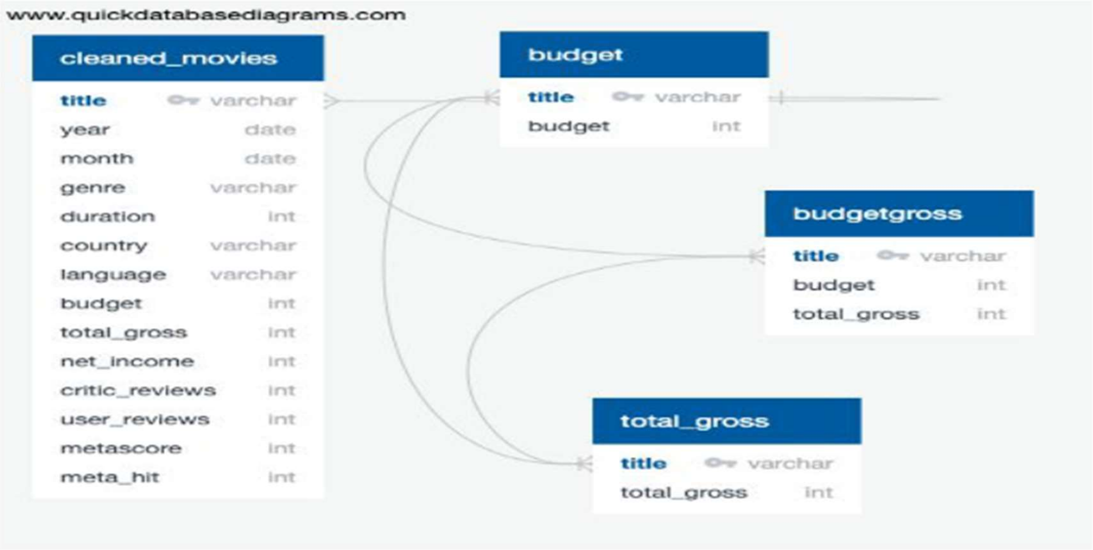
In [22]:
```
report = classification_report(y_test, y_pred)
print(report)
```

```
              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
```

•SQLite database was created using DB Browser for SQLite
•Database were stores and data were using in the project
•Database was connected to the model by SQAlchemy scripts
•Includes 4 tables, one of which was created by joining two tables together
•ERD, SQL join code, and database structure pics attached

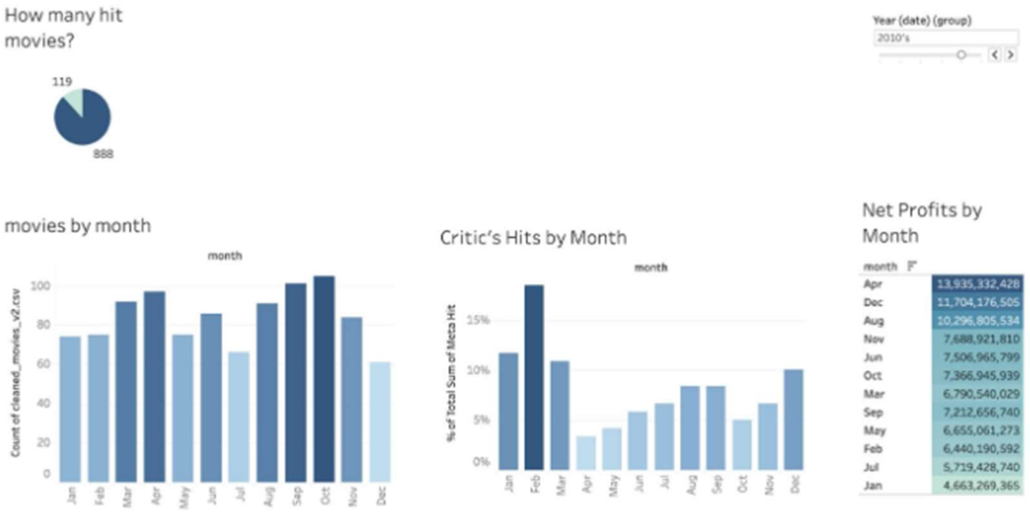


cleaned_movies	
title	varchar
year	date
month	date
genre	varchar
duration	int
country	varchar
language	varchar
budget	int
total_gross	int
net_income	int
critic_reviews	int
user_reviews	int
metascore	int
meta_hit	int

budget	
title	varchar
budget	int

budgetgross	
title	varchar
budget	int
total_gross	int

total_gross	
title	varchar
total_gross	int



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