**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.

![A picture containing graphical user interface

Description automatically generated]()

* 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

Graphical user interface, application

Description automatically generated

Chart, bar chart

Description automatically generated