**Project Title:** The higher the budget the more gross a movie produces.

**Team Members:** Amandeep Brar | Arpine Bankikyan | Cassie Folkers | Lindsey Giron | Ricardo Negrete | Sriven Ankam

**Project Description/Outline:**

* Reviewing all genres
* Profit recognition
* Awards won
* Budget to gross
* 5 years range

**Research Questions to Answer:**

1. Does a higher budget produce a more profitable movie?
2. What genres yield the biggest bang for the buck based on budget to gross?
3. Does a larger budget yield a higher popularity?

**Datasets to Be Used:**

* The Movie Database API –
* IMDb
* OMDb
* Kaggle – Netflix regarding original titles

**Rough Breakdown of Task:**

Documentation – everyone will keep track and contribute

Identify datasets to be used and assure it contains the data we need for analysis – this was done as a group in class

Pull dataset – lead by Cass (Each team member will pull a year from API)

Clean datasets – lead by Sriven – this was done during class

Create script to extract data for analysis – Sriven/Ricardo

Outliners, linear regression, correlation – Lindsey

Create plots for visualization – Amandeep

Analysis – Apri

Presentation – Ricardo

Ultimately assist each other as a group!

**NOTES**

\*\*\* The Movie DataBase Statement:  
“You shall use the TMDb log to identify your use of the TMDb’s API’s.

You shall place the following notice prominently on your application: ‘This product uses the TMDb API but is not endorsed or certified by TMDb.’

Any us of the TMDb log in your application shall be less prominent than the logo or mark that primarily describes the application and your use of the TMDb logo shall not imply any endorsement by TMDb,”\*\*\*

\*\*\*OMDb

* TMDb publicly maintained but only database with budget and revenue data
  + Created profit data
  + Budget is defined
  + Revenue is defined
* OMDb was used to pull metacores, rating and votes
  + Needed to remove OMDb due to little data from metascores 9043 out 84294
* IMDb was used to get movie titles
* Netflix - ratings
  + 568 titles we were only data available 44 titles.
* Cleaning data
  + Remove titles missing data for budget, revenue and genre
  + Needed to fill missing data with NaN to use .dropna
* Merge clean data
* Save pulled data into csv files
* Create visualizations
  + Identify any outliers
    - Remove upper and lower outliers
    - Largest outlier was Secret Superstar was over 40,000% from the upper quartile
  + Create scatter plots with linear regression
  + The top 3 most genres being made are Drama, Action, and Comedy

Our hypothesis was that having a larger budget yields a larger profit. Our reasoning is that you are able to have a better director, actor/actress, location/s, effects and marketing. The first limitation was to find a database with budget and revenue data as there are non-disclosure agreements for some films. This did not stop our curiosity. We found The Movie Database (TMDb) which is publicly maintained and obtained budget and revenue data which was crucial to answer for our hypothesis. The IMDb had movie titles and ratings. OMDb had metascores and awards. Netflix has rating scores for their original titles to take a look at streaming data.

During the pulling process we had to split the TMDb data by year since we were puling a large amount []

Each team member was assigned a year to pull after the script was created. Once we had all the data pulled, we worked together in cleaning and merging the data. Once the data was merged and missing data was removed, we discovered that would not be able to use Netflix as they only had the full data for 44 titles out of 568 titles. Also, OBMd only had 9043 titles with the full data our of 84,294 titles.

TMDb had budget and revenue data which we were able to create the profit data for each movie. We decided to create scatter plots with linear regressions since we wanted to know if there was a strong correlation between budget and profit.

The Budget vs Profit scatter plot displayed no correlation with a r-squared result of 0.04, instead it showed that a high profit can be produced with a lower budget. Most of the high profits were seen in budgets under 50 million, granted the majority of the movies fit below the 50 million budget range. While some movies with higher budgets did produce a high profit with fewer below zero as we get closer to a higher budget, this is also achieved with a lower budget.

The next question is out of the top three genres is there a correction between budget vs profit. Drama, Action, and Comedy were our top three genres. Drama movies showed no correlation between budget to profit. It also had the majority of its films below a 50 million budget, but the majority of its profits were under a 20 million budget. Action movies scatter plot was a bit more spread out as this is a genre which clearly requires a larger profit for effects which as we can see do yield high profits but we also see high profits and more with movie budgets under 100 million. Therefore, no correlation with a r-squared result of 0.13. Comedy similar to Drama have the majority of their budgets on the lower end, specifically under 60 million and we see those movies yield over 100 profits.

The top tree genres are drama, action, and comedy.