***Technical Report***

***Sources of data that you will extract from:***

1. CSV files predominantly from Kaggle. We were fortunate enough to have many options on movie-related data.

***The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc.)***

1. Cleaning – Reviewing columns to ensure that data entered is what is expected, i.e., in some columns where it should be there year produced, there were strings.
2. Filtering – For the sake of filtering, we selected only movies produced in the U.S, between 2000 and 2017.

***The type of final production database to load the data into***

1. SQL– as our data is fairly structured in each dataset, with one unique movie on each row with multiple data points, we elected to put our data in to SQL.

***The final tables or collections that will be used in the production database.***

1. IMBD Data – Data from an IMBD dataset that includes movie title, year, genre, duration (length), director, description, average vote, number of reviews from users, and number of reviews from critics. We selected this dataset as the review columns could be interesting in analysis.
2. Netflix Data – Data from a Netflix dataset that includes, movie title, director, actors, year, rating, duration (length), description. We chose this data set for the details including the description, which could be valuable for a scraping analysis.
3. Oscar wins – Data from Oscar winning films including movie title, year, and if the film won best picture, best director, best actor, best actress, best supporting actor, and best supporting actress. We selected this dataset as there could be interesting analysis done on which films won awards, using data from other datasets to perform statistical analysis.
4. Oscar Data – Data from Oscars academy including movie title, certificate (parental rating), duration, genre, Oscar rating, number of votes, count of user review, count of critic review, and if the movie was a cartoon. We selected this database to corroborate with the Oscar wins database, gathering more datapoints on the voting process. The Boolean field for cartoon could also be used to perform additional analysis.
5. Wiki Data – Data from Wikipedia including movie title, year, country, genre, and Wikipedia page. We selected this dataset as the Wikipedia page links could be utilized to scrape and pull-down additional data points.