This project is divided into two parts:

* **The Story of Film:** This section aims at narrating the history, trivia and facts behind the world of cinema through the lens of data. Extensive Exploratory Data Analysis is performed on Movie Metadata about Movie Revenues, Casts, Crews, Budgets, etc. through the years. Two predictive models are built to predict movie revenues and movie success. Through these models, we also aim at discovering what features have the most significant impact in determining revenue and success.
* **Movie Recommender Systems:** This part is focused around building various kinds of recommendation engines; namely the Simple Generic Recommender, the Content Based Filter and the User Based Collaborative Filter. The performance of the systems are evaluated in both a qualitative and quantitative manner.

**Approach**

The problem was divided into several steps:

1. **Data Collection:** Data was collected from the MovieLens website and through a script that queried for data from various TMDB Endpoints.
2. **Data Wrangling:** The datasets were uploaded to a dataframe and explored. Null values were filled in wherever appropriate and polluted values were discarded or wrangled.
3. **EDA:** Extensive data visualisation and summary statistics were used to extract insights and pattern from the various datasets. The history, facts and trivia behind movies were narrated through data.
4. **Machine Learning:** Gradient Boosting Classifer and Regressor were trained on our feature engineered dataset to predict movie success and revenue respectively. Their feature importances were noted to gain insights into what factors influence the revenues of a movie relative to budget.
5. **Recommendation Systems:** Four different recommendation systems were built using various ideas and algorithms such as IMDB's Weighted Rating, Content Based Filtering and Collaborative Filtering.

**Final Results**

A Gradient Boosting Regressor and Classifier were built to predict Movie Revenue and Success respectively with a Score of 0.78 and 0.8 respectively.

In addition, four recommendation engines were built based on different ideas and algorithms:

* **Simple Recommender:** This system used overall TMDB Vote Count and Vote Averages to build Top Movies Charts, in general and for a specific genre. The IMDB Weighted Rating System was used to calculate ratings on which the sorting was finally performed.
* **Content Based Recommender:** I built two content based engines; one that took movie overview and taglines as input and the other which took metadata such as cast, crew, genre and keywords to come up with predictions. I also devised a simple filter to give greater preference to movies with more votes and higher ratings.
* **Collaborative Filtering:** I used the powerful Surprise Library to build a collaborative filter based on singular value decomposition. The RMSE obtained was less than 1 and the engine gave estimated ratings for a given user and movie.
* **Hybrid Engine:** I brought together ideas from content and collaborative filtering to build an engine that gave movie suggestions to a particular user based on the estimated ratings that it had internally calculated for that user.