How to make money in Hollywood

A Project Proposal

CMPG-764 Assignment #3 Phase I

By Michael Rozycki, Steven Sikorsky, Muhammad Buttar

## Table of Contents

[**Table of Contents**](#_apourwepr74p) **2**

[**[1] Goal and Problem Statement**](#_sf7oaycknh6i) **3**

[**[2] Background**](#_71ajff7a722e) **3**

[**[3] Methodology**](#_o5nhgro251wi) **4**

[**[4] Information about the dataset**](#_iyod96wpobec) **5**

[**[5] Project Timeline: Total estimated time to complete**](#_17361ddmb7i9) **5**

## 

## [1] Goal and Problem Statement

Preliminary knowledge about films such as *Night of the Living Dead* and *Avengers: Endgame* suggests that the horror genre is more profitable with respect to the amount invested, amount earned ratio, while the action genre provides the opportunity to earn large quantities of money faster but at the cost of being less economical to start with. In this project, we seek to determine whether these preliminary inferences hold throughout a genre when the outliers are not the sole source of assessment. To do this, we will attempt to answer the questions: Does a specific genre of movies provide a better return on investment ratio? Is it better to produce a film with a respectively low risk (that is, a small investment) or to produce a film with a respectively high risk (that is, a large investment)?

From these questions, as well as others, we will determine which genre of films has historically the best return on investment.

## [2] Background

There are many datasets publicly available containing data on film releases throughout the decades. These datasets contain data such as the name of a film, its genre, its year of release, its budget, and its box office earnings.

We may obtain datasets that are available for us to use from potential sources such as IMDb, The Numbers, and Box Office Mojo (in addition to many others).

Additional questions that may be answered when analyzing the data are:

* + Has time affected the return on investment ratios seen by each genre, or have they remained consistent with time?
  + How risky are big budget films, do some genres have more flops than others?
  + If data is available on the average rating of each film, then how might ratings be involved? Does the rating correlate with the return on investment?
  + If budget components are available, what insights can be gained from this? Is a certain proportion of the budget necessary for a particular genre, and is it different by genre?

## 

## [3] Methodology

Investigation into potential datasets has shown the potential to either obtain a dataset or compile a dataset from partial datasets that will have entries containing the following attributes: the year of release, the title, the genre, the budget, and the box office earning. We will take data from the past few decades containing this information, and balance the dataset to contain equal representation of each genre.

To make use of the available data, we will be focusing on the computation of the various ratios that can be derived from the budget and box office earnings. We will go by genre and calculate the ratios between each film’s budget and its total box office earnings. From this we can find the average ratio and median ratio for each genre. Additional analyses will also be done.

For Hadoop Map-Reduce, the work to be done by each phase is as follows:

**The Mapper**

* Generates Key-Value (K-V) pairs
  + Genre:Ratio(s)
    - Ratio between budget (individual) and box office (individual)
    - Ratio between budget (production) and budget (advertising) \*if data available\*
  + Year:Genre:Ratio(s)
  + Year:Title:Genre:Ratio(s)
  + Genre:Budget
  + Genre:BoxOffice

**The Reducer**

* Performs aggregation operations for each key
  + Average the K-V pairs
  + Count # of pairs above / below certain thresholds
  + Find Maximum ratio for a genre (from the Genre:Ratio K-V pairs, and from the Y:T:G:R K-V pairs, where a film can be uniquely identified later using the year and title, and then pertaining to its genre)
  + Find the median value of a Key
  + Find the ratio between averaged budget and averaged box office for each genre

## [4] Information about the dataset

In terms of the dataset, there are a few different datasets we have found online that could potentially be useful. Some are downloadable files, some consist of data on websites. To retrieve the specific data we need for our project’s purposes, we will likely need to create our own dataset derived from information from the previously mentioned potential datasets. We expect this to take approximately 1 week to collect and prepare this data to have what we need.

For the preparation of our dataset, we will need to combine various data into a single dataset containing: the year of release, title, genre, budget, and box office earnings. We will include a breakdown of each budget if the data is available for a large enough quantity of entries. The compilation will be done by applying joining techniques with the title of a film and its release year as the primary key. An example of what an entry in our dataset will look like is:

Year Title Genre Budget Box Office

1984 “The Terminator” Action 6400000 78680331

Where the budget and box office values are measured in USD.

## [5] Project Timeline: Total estimated time to complete

The following is our estimation of the project’s timeline:

Phase 2: Collect data, prepare, sanitize, and clean if required: 1 week

Phase 3: Implement your Hadoop Map-Reduce Code: 1 week

Phase 4: Run map-reduce and collect results 1 week

Phase 5: Analyze and present your results: 1 week

