IMDB未定

**Abstract**

Collecting data from IMDB, formerly known as Internet Movie Database, we get a dataset about 5043 movies crossing 66 countries over 100 years. After Data cleansing, we use this dataset to build a general linear model to figure out what contribute to the movie reputation and its gross.

我们得出的结论12345

1. **Introduction**

**A description of**

1. **the relevant background of the topic**

How can we judge a movie while it is even unreleased?

This simple question puzzled us for a very long time since there is no universal way to claim how good is the movie. Many people rely on the critics to gauge the quality of a film, while others use their instincts. But It take time to obtain a reasonable amount of critics review after a movie is released. Even if we collect enough reviews, human instinct sometimes is unreliable.

1. **the motivation of the project.**

Movie is considered to be an important art form and a source of popular entertainment. Considering the value of movie, it is necessary for us to find out how the basic elements contribute to the reputation or the gross.

1. **questions of interest**

We want to simply judge whether a movie is great by its gross and the IMDB score. The questions we are considered is list below

1. What contributes to the IMDB score or the gross of a movie?
2. What are the most important variables?
3. The relationship between IMDB scores and the gross.
4. **Data description**

The dataset is scraped from IMDB website by using a Python library called “scrapy”.

There are 28 variables for 5043 movies and 4906 posters (998MB), spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses.

Those 28 variables include: the title of movie, colored or not, released year, duration, language, country, content rating, genres, budget, gross, movie Facebook likes, director’s name, director’s Facebook likes, 3 of the main actors’ name, their Facebook likes, total cast Facebook likes, number of voted users, aspect of ratio, the IMDB score, the number of critic for reviews, the face number in the poster, plot keywords, IMDB link.

During the progress of modeling, we will drop all the names and use some of the variables listed above.

1. **Methods and Results:**

首先，扔了很多数据，有一些是因为他们不是可量化的指标，比如说电影的名字，一些是level太多，比如说具体是哪位演员，关键词，这些指标样本数太少。

其次，我们对关键的一些变量做了一些正态检验，看看误差什么的。

然后我们对数据进行建模

找出显著的

然后再F test一下这些不显著的

然后再进行下一步建模

Blablabla

1. **Conclusions and Discussion**

特别大的影响因素（b很大的，而且显著的）

显著但是影响程度不大的因素

意味会显著，但并没有的因素

1. **Appendices**

这里放输出结果图

1. **References**

**不知道！**