# 项目：可视化电影数据

## 第一步：清理数据和选择变量

- 清理数据：

  1、将数据文件读入 python，删除杂行数据

  ```
  movies = movies[movies.iloc[:,21].isnull() & movies.iloc[:,22].isnull()]
  ```

  2、去掉重复行

  ```
  movies = movies.drop_duplicates()
  movies.to_csv('movies1.csv')
  ```

  3、在表格中通过 release_year 筛选非年份的数据行，去掉错乱行

  4、通过 excel 分列功能将 genres 列分列

- 选择变量：

  根据需要解答的问题，选定了'id', 'original_title', 'keywords', 'genres', 'production_companies', 'release_year', 'vote_count', 'vote_average', 'budget_adj', 'revenue_adj' 这些变量，删除其他列，将数据文件导入 Tableau。

## 第二步：问题

- 回答下列问题，引用你在线可视化结果去支持你的答案：
  - **问题 1：**电影类型是如何随着时代变化而变化的?
  - **问题 2：**环球影业和派拉蒙影业的电影之间数据指标有什么区别?
  - **问题 3：**和非小说改编的电影相比，基于小说改编的电影表现得怎么样?
  - **问题 4：叫好又叫座的电影与叫好不叫座的电影有什么区别？**
- 你提出的另外问题是什么？答案是什么？你是怎么想出这个问题的？

  我提出叫好不叫座的电影与叫好又叫座的电影有什么区别，是因为想知道有些电影能够得到高评分，票房却并不乐观的原因。

  从可视化图中可以看出：

  1. 评分人数与票房有一定联系，票房不高，很可能是因为传播度不够。
  2. 题材与票房也有一定关系，像冒险、动作、科幻等题材就非常受影院用户的欢迎。

  当然，还有很多因素可能影响电影叫好却不叫座，譬如演员人气、前期宣发，这部分还需要其他数据补充，才能进一步研究。

## 第三步：可视化

链接：
https://public.tableau.com/views/movies2/Q4?:embed=y&:display_count=yes&publish=yes