项目：可视化电影数据

第一步：清理数据和选择变量

- 清理缺失的信息，选择在可视化中需要进一步研究的最重要的变量。
- 列出你在可视化中会进一步研究的变量。你要研究的变量数量应该不超过 8 个。

清理数据：

1. 在 excel 中，对 movies.csv 的 genres 变量进行分列，分成 5 列。
2. 在 tableau 中，对 genres 五列进行数据透视表，生成一个新的变量 genres_new
3. 对 genres_new 的缺失值进行筛选就可以，genres_new 变量做筛选器，把 null 的值排除
4. 对 production_companies 变量分组，形成新变量 production_companies (组)，这个变量包含 Paramount Pictures、Universal Pictures 和其他。
5. revenue_adj，budget_adj 波动较大，此项目采取了移动平均值。
6. keywords 变量分组，形成新的变量 keywords（组）。这个变量包含改编电影、原创电影。分组的依据是值是否含有 based on novel

研究变量：

个数	变量名称
1	genres_new
2	production_companies (组)
3	release_year
4	revenue_adj
5	budget_adj
6	keywords（组）
7	popularity

第二步：Tableau 可视化

工作簿
地址：
https://public.tableau.com/views/P3_movies_3/Q1?:embed=y&:display_count=yes&publish=yes

重要提示: 请把你的工作簿上传到 **Tableau Public** 上，并允许你的项目评审师查看你的工作簿。请注意单单把文件简单保存为后缀名为 “.twbx” 不能让所有项目评审师查看文件。如何做到能够允许所有项目评审师查看文件的指示在这里。

第三步：问题

问题 1：电影类型是如何随着时代变化而变化的?（story）

80 年代之前，各个电影类型的数量差距并不是很大，但是从 1990 年之后开始，每年各个电影类型出现的个数出现明显的差异。Drama 类型的电影随着时间的变化，个数极速增长。2015 年时数量已经接近 300 个。而 Western、TV Movie、War 等类型的电影从 1960 年开始到 2015 年没有怎么变化，每年此类型的电影数量都是相对较少的。

Drama，Comedy，Thriller，Action，Romance，Horror，Adventure 这几类电影是最长被制作的。

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?

从电影类型来说，这两个制作公司的电影类型主要偏向 Comedy，Drama，Action，Thriller 这几个类型。其中 Paramount Pictures 更喜欢制作 Drama，而 Universal Pictrues 更喜欢制作 Comedy 类型电影.

从 popularity 和 vote_average 来讲，Universal Pictrues 比 Paramount Pictures 表现的要好。

从 90 年代后期开始，两个公司的预算开始超过收入。2010 年之后，Paramount Picture 的 revenue_adj 比 budget_adj 高一些了，而 Universal Pictures 的在 2012 年后 revenue_adj 拉开了与 budget_adj 的距离，收入大于预算。

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?

小说改编的电影个数远远低于非小说改编的电影个数。

小说改编的电影类型主要集中在 Drama，Comedy，像 Western，Histtory 等电影类型是没有小说改编的。

小说改编电影的评分会相对于非小说改编电影高。

问题 4：revenue_adj 与 popularity 的关系？ 你是怎么想出这个问题的？

revenue_adj 与 popularity 呈现一个正相关，其中 Adventrue 的收入越高，其 popularity 的比例越高。而 Documentary、TV Movie 这几类电影可能会出现 revenue 不高，但是 popularity 比较高。

这个问题的提出，有助于公司了解 revenue_adj 与 popularity 的关系，以及到底哪类电影最容易出现 revenue 与 popularity 都高的可能，并且