# 练习题： 观影大数据分析

王S聪想要在海外开拓万D电影的市场，这次他在考虑：怎么拍商业电影才能赚钱？毕竟一些制作成本超过1亿美元的大型电影也会失败。这个问题对电影业来说比以往任何时候都更加重要。所以，他就请来了你们队来帮他解决问题， 给出一些建议， 根据数据分析一下商业电影的成功是否存在统一公式？以帮助他更好地进行决策。

解决的终极问题是： **电影票房的影响因素有哪些？**

接下来我们就分不同的维度分析：

- 观众喜欢什么电影类型？有什么主题关键词？
- 电影风格随时间是如何变化的？
- 电影预算高低是否影响票房？
- 高票房或者高评分的导演有哪些？
- 电影的发行时间最好选在啥时候？
- 拍原创电影好还是改编电影好？

本次使用的数据来自于 Kaggle 平台（TMDb 5000 Movie Database）。收录了美国地区 1916-2017 年近 5000 部电影的数据，包含预算、导演、票房、电影评分等信息。原始数据集包含 2 个文件：

- tmdb_5000_movies：电影基本信息，包含 20 个变量
- tmdb_5000_credits：演职员信息，包含 4 个变量

请你们队完成下列问题：

（1）使用附件中的 tmdb_5000_movies.csv 和 tmdb_5000_credits.csv 数据集，进行数据清洗、 数据挖掘、数据分析和数据可视化等，研究电影票房的影响因素有哪些？ 从不同的维度分析电影，讨论并分析你的结果。

（2）附件 tmdb_1000_predict.csv 中包含 1000 部电影的基本信息， 请你选择合适的指标，进行特征提取，建立机器学习的预测模型， 预测1000部电影的 vote_average 和 vote_count，并保存为 tmdb_1000_predicted.csv。

（3）从数据的分析与研究中，你是否能找到一个确定拍摄一部电影最有可能赚到钱的模式或决策方法？如果能，请你给出你的方法并说明它的理由。如果不能，请你也给出你的理由与依据。