项目：探索影响电影票房利润的各类因素

第一步：问题提出

1，预算与净利润之间有什么关系？

2，针对净利润一项，平均利润最高的 5 位的演员是谁？

3，针对净利润一项，平均利润最高的 5 电影类别是哪些？

4，针对净利润一项，平均利润最高的 5 个上映月份是哪几个月？

第二步：数据整理

数据筛选和补充（工具：jupyter notebook）

① **删除问题中与问题不相关的列**

② **新增利润一列：profit**

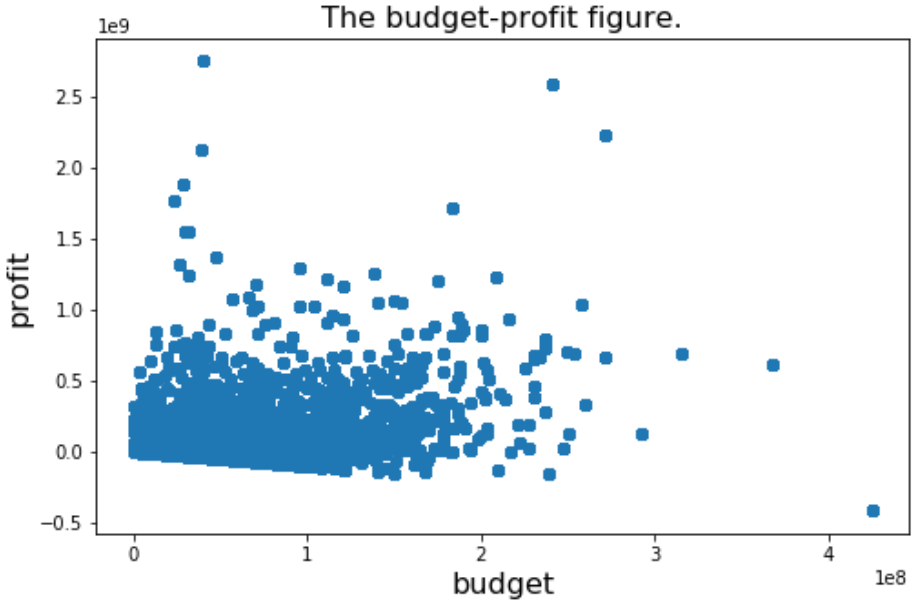
③ **将"cast"、"genres"两列按照"|"进行拆分和重组**

④ **去除重复行**

⑤ **针对各项问题的解决，复制和处理生成新的 dataframe**

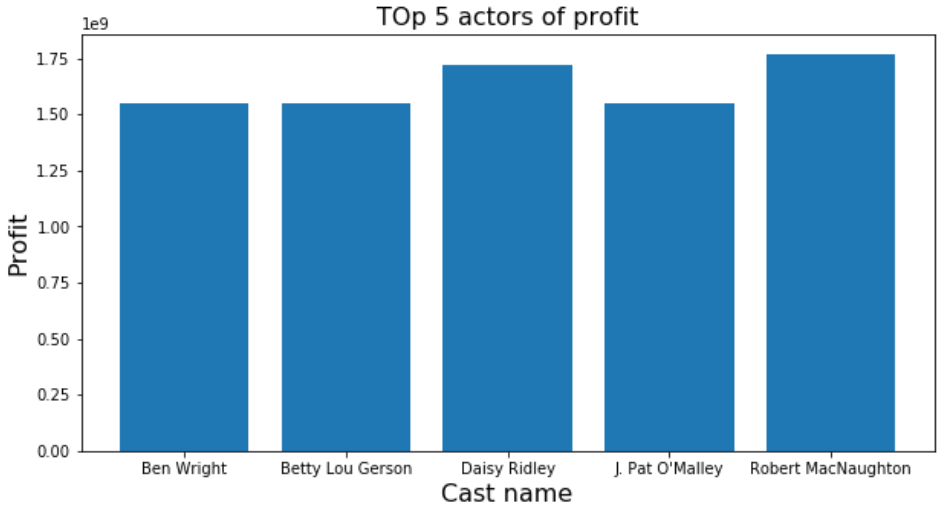
第三步，数据分析和可视化

1，预算与净利润之间有什么关系？



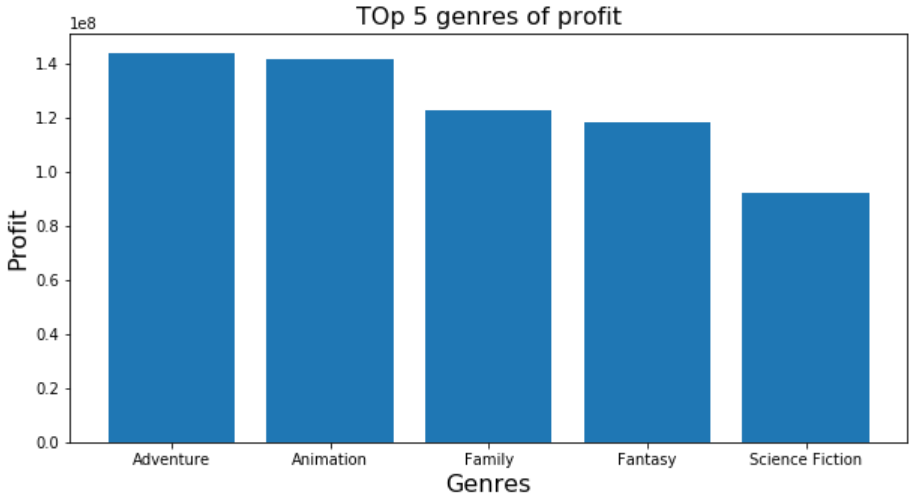
通过计算和绘制散点图如上，净利润与预算之间的相关性为正相关，皮尔逊相关系数为 0.42，为中等相关关系。

2，针对净利润一项，平均利润最高的 5 位的演员是谁？



以电影平均的净利润值为标准，获得净利润最高的前 5 位演员分别是：Robert MacNaughton、Daisy Ridley、Betty Lou Gerson、J. Pat O'Malley、Martha Wentworth，各演员对应的净利润均值分别为 17.68 亿美元，17.19 亿美元，15.46 亿美元，15.46 亿美元，15.46 亿美元，分别对应上方条形图所示。

3，针对净利润一项，平均利润最高的 5 电影类别是哪些？



通过计算，平均利润最高的 5 电影类别分别是：Adventure（冒险类）、Animation（动画类）、Family（家庭类）、Fantasy（幻想类）、Science Fiction(科幻小说类) ,对比如以上条形图所示 ,其对应的平均票房利润分别是 1.44 亿美元、1.42 亿美元、1.23 亿美元、1.18 亿美元、0.92 亿美元，可以预期这五类电影较受欢迎。

4，针对净利润一项，平均利润最高的 5 个上映月份是哪几个月？

针对一年中 12 个月里的电影票房净利润探索，平均值最高的 5 个月份分别为：6 月、12 月、11 月、5 月、7 月，对应值分别为 1.20 亿美元、0.98 亿美元、0.91 亿美元、0.89 亿美元、0.84 亿美元，如上图所示。

第四步，得出结论

1，结论总结

Q1：预算与净利润之间有什么关系？

A1：并非很强的线性关系，对于预算高的电影，会有相对较高的利润预期，但并没有十分明显的指导意义。

Q2：针对净利润一项，平均利润最高的 5 位的演员是谁？

A2：分别是 Robert MacNaughton、Daisy Ridley、Betty Lou Gerson、J. Pat O'Malley、Martha Wentworth，此 5 位影星或较受观众喜爱，对于电影演员列表中含有此 5 位演员的电影，可能会有高利润预期。

Q3：针对净利润一项，平均利润最高的 5 电影类别是哪些？

A3：Animation（动画类）、Adventure（冒险类）、Family（家庭类）、Fantasy（幻想类）、Science Fiction(科幻小说类)，此 5 类电影或较受观众青睐。

Q4：针对净利润一项，平均利润最高的 5 个上映月份是哪几个月？

A4：6 月、12 月、5 月、11 月、7 月，该 5 个月份对应寒暑假、圣诞节期间，在这几个月份中上映的电影，或有较高的票房预期。

2，遇到的问题及说明

① 本项目在分析过程中，每一个分析结论的得出，均基于较大的数据规模（5165~30525），数据规样本规模足够。

② 本项目原始数据罗列了 10 余个特征，对于票房分析来讲，已经足够全面涵盖各种因素，足以对探索目标进行分析和归纳，但本项目探索中，仅针对其中的演员、上映月份、影片类型、预算及利润这 5 个特征进行分析总结，并没有对所有特征进行归纳，所以得出的结论仅基于影响电影商业价值的部分要素。

③ 本项目在数据清洗的过程中，主要去除了所分析列中为空值的数据组，这部分数据来源来自数据缺失，即无法利用该数据组进行分析，在剔除这部分数据后，虽然仍有较大规模的数据，且分析结果仍然较为可信，但无法避免因数据缺失导致的结论偏差。

④ 本项目针对 5 个特征的归纳仅为初步探索，尤其对演员的探索中，仅仅输出票房最高的前五位演员，而未对每个演员出演的电影个数，即出镜率进行统计，有可能存在出镜率较低 ,但其中某几部电影的票房异常高而拉高整体平均票房的情况。且由于演员人数众多，或许研究 Top50 的演员，对于指导哪些演员出演的电影更受欢迎，票房更高具有更现实的意义。出于绘图和篇幅限制，仅对此 Top5 的演员的平均票房的计算，难免会存在误差。

⑤ 本项目中对 5 个特征进行探索，且 4 个结论均只基于单个特征进行归纳，并未总体考虑进行总结 ,可能存在分析错误 ,如 :最受欢迎的电影类型为 Adventure（冒险类），且此类别电影恰好集中在 6 月份上映较多，导致关于最佳上映月份的探索结果中，误以为所有类型的电影均在 6 月份上映有较好的票房预期。又如 : 票房最高的几个演员之所以拥有高票房 , 可能是由于其出演电影的导演水平高，或者较受观众欢迎，而并非因为演员演技好或者粉丝多而获得高票房。

⑥ 本项目未进行统计学显著性检验，也未经过算法计算，而是仅将数据部分初步可视化，所以所得结论仅为初步结论，结论仅为特征的归纳，为获得高票房的必要条件，但未必是充分条件，即尚无法得出严密的因果关系结论。所以本项目结论仅有初步参考意义，还需要更深入的综合研究，还有进一步探索的空间。

崔传敏

20190724