MathorCup 高校数学建模挑战赛——大数据竞赛

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（数据分析师）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

请使用 Python 编程，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。

数据清洗

1 导入数据

2 缺失值处理

缺失记录仅_____条，采取网上搜索，补全信息。

2.1 补全 release_date

缺失记录的电影标题为《_____》，日期为_____。

2.2 补全 runtime

缺失记录的电影 runtime 分别为_____min 和 _____min。

3 重复值处理

运行结果：有_____个不重复的 id，可以认为没有重复数据。

4 日期值处理

将 release_date 列转换为日期类型：

5 筛选数据

使用数据分析师最喜欢的一个语法：

票房、预算、受欢迎程度、评分为_____的数据应该去除；

评分人数过低的电影，评分不具有统计意义，筛选评分人数大于_____的数据。

此时剩余_____条数据，包含_____个字段。

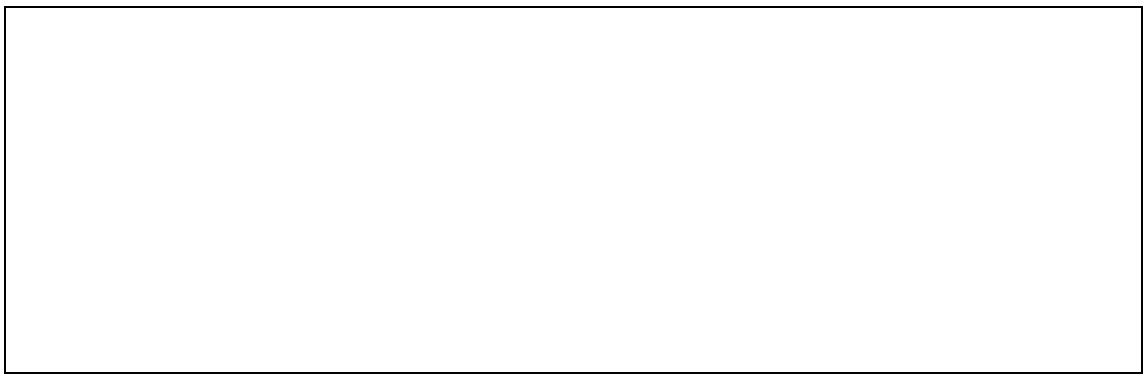
6 json 数据转换

说明：genres,keywords,production_companies,production_countries,cast,crew 这 6 列都是 json 数据，需要处理为列表进行分析。

处理方法：

json 本身为字符串类型，先转换为字典列表，再将字典列表转换为，以’,'分割的字符串

7 数据备份



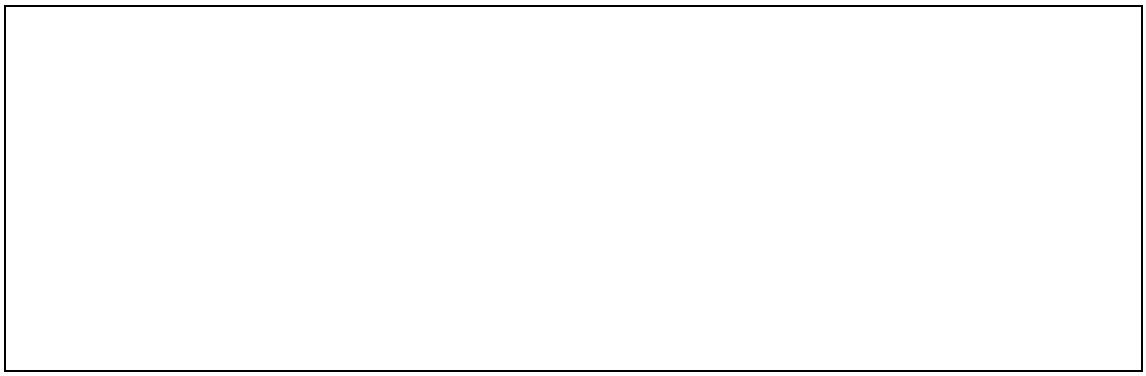
5 数据分析

5.1 why

想要探索影响票房的因素，从电影市场趋势，观众喜好类型，电影导演，发行时间，评分与关键词等维度着手，给从业者提供合适的建议。

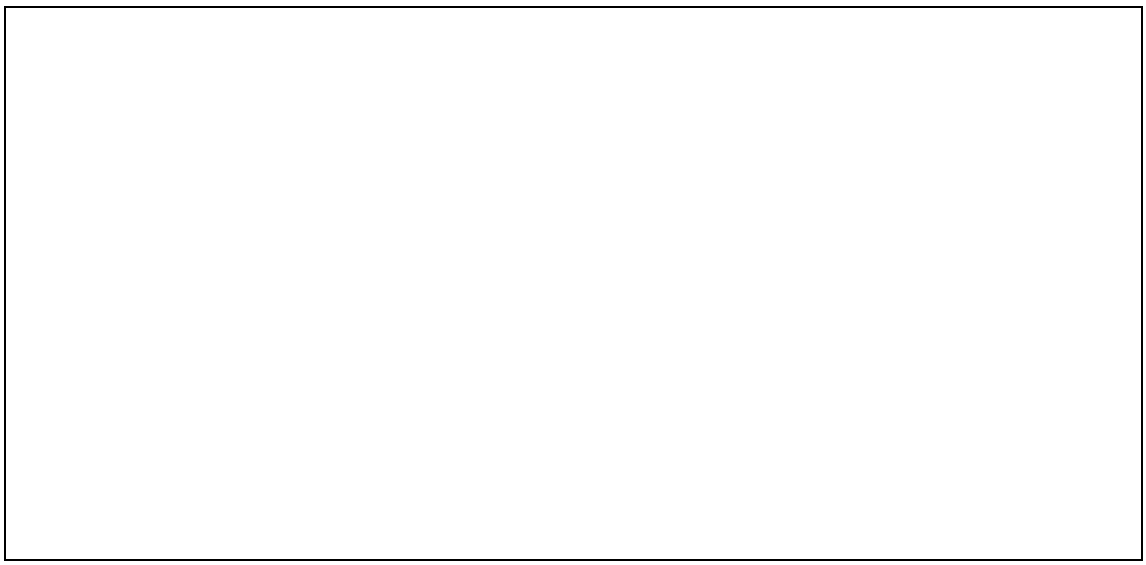
5.2 what

5.2.1 电影类型：定义一个集合，获取所有的电影类型

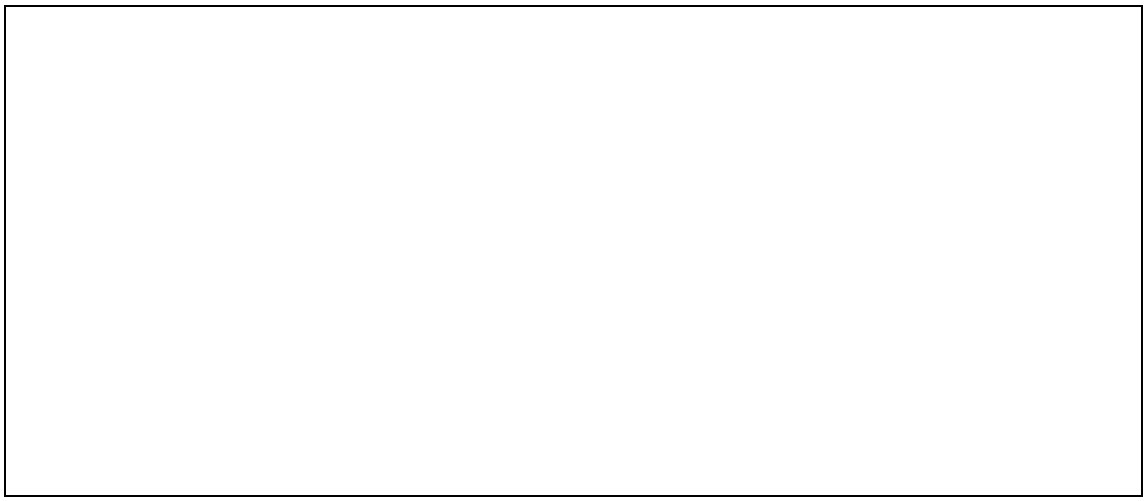


注意到集合中存在多余的元素：空的单引号，所以需要去除。

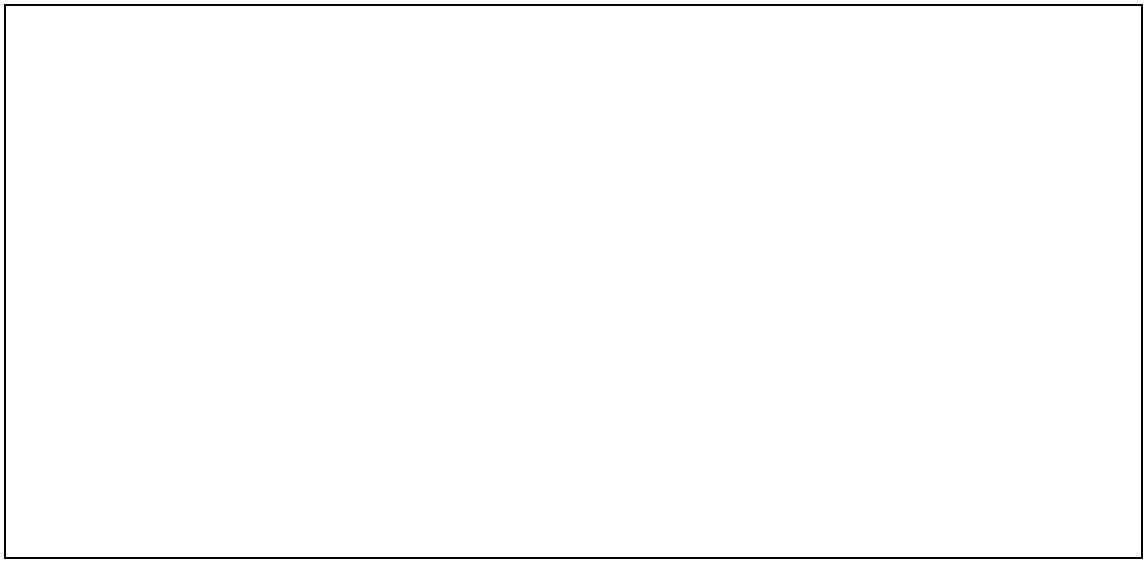
5.2.1.1 电影类型数量（绘制条形图）



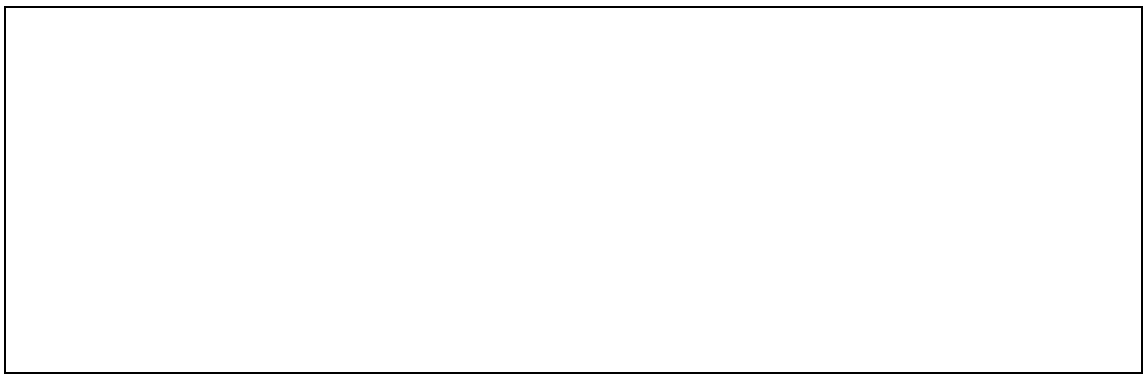
5.2.1.2 电影类型占比（绘制饼图）



5.2.1.3 电影类型变化趋势（绘制折线图）



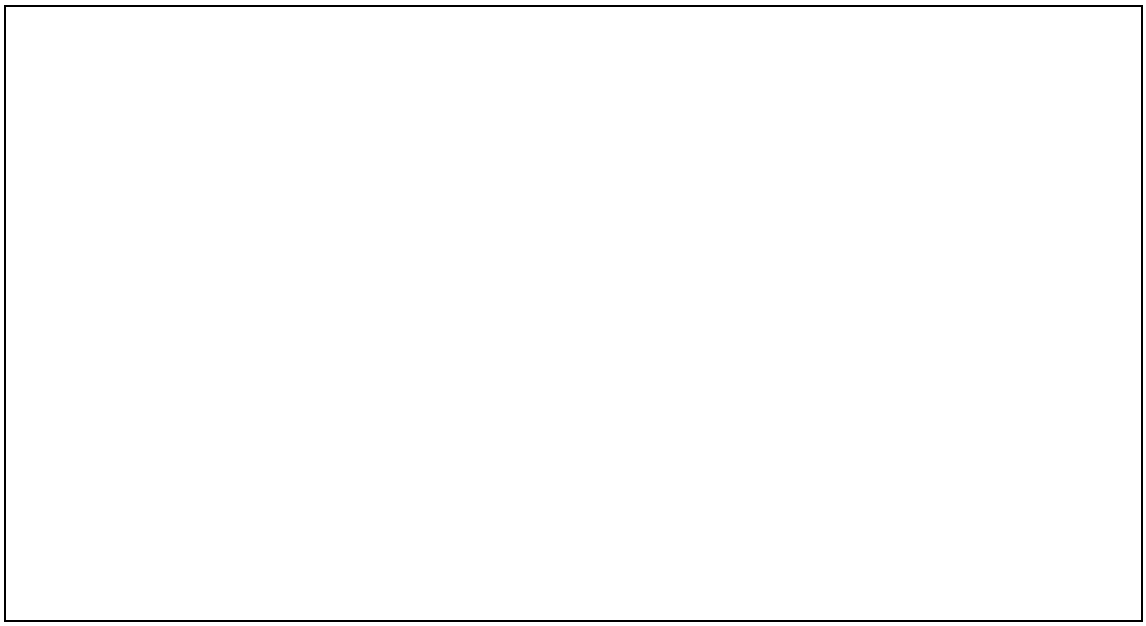
5.2.1.4 不同电影类型预算/利润（绘制组合图）



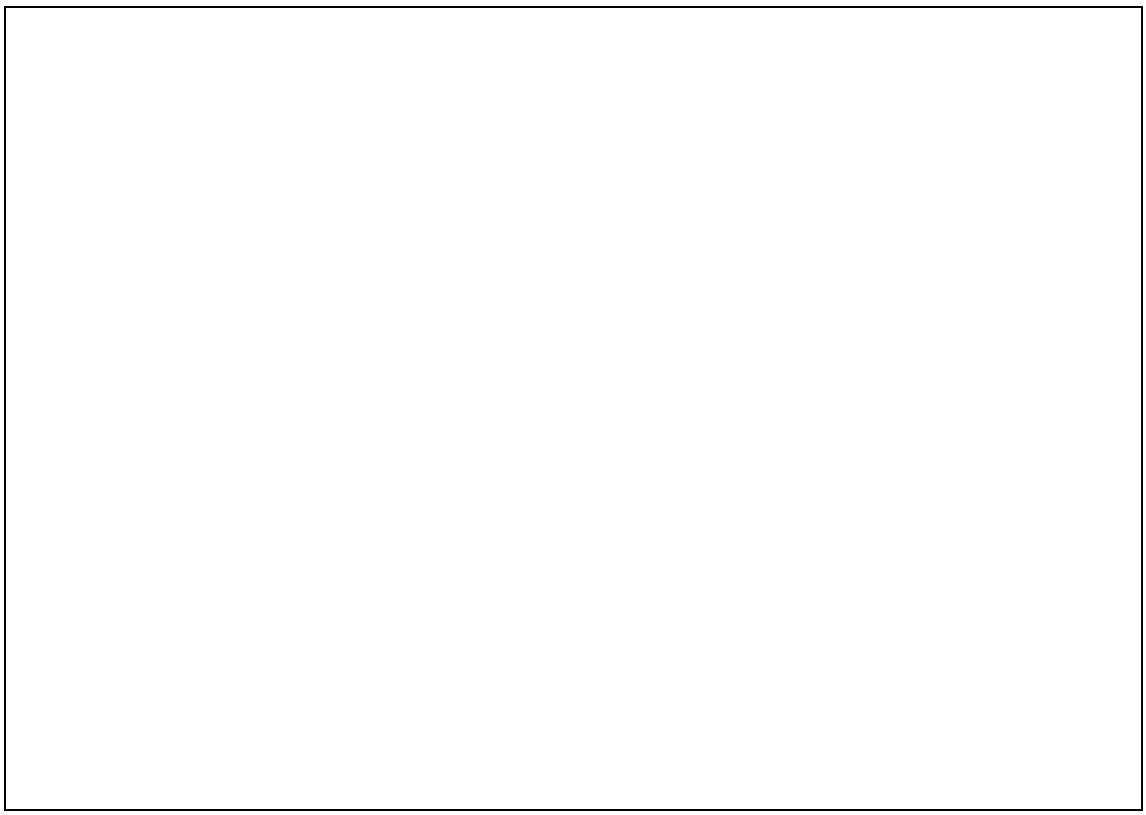
5.2.2 电影关键词（keywords 关键词分析，绘制词云图）

5.3 when

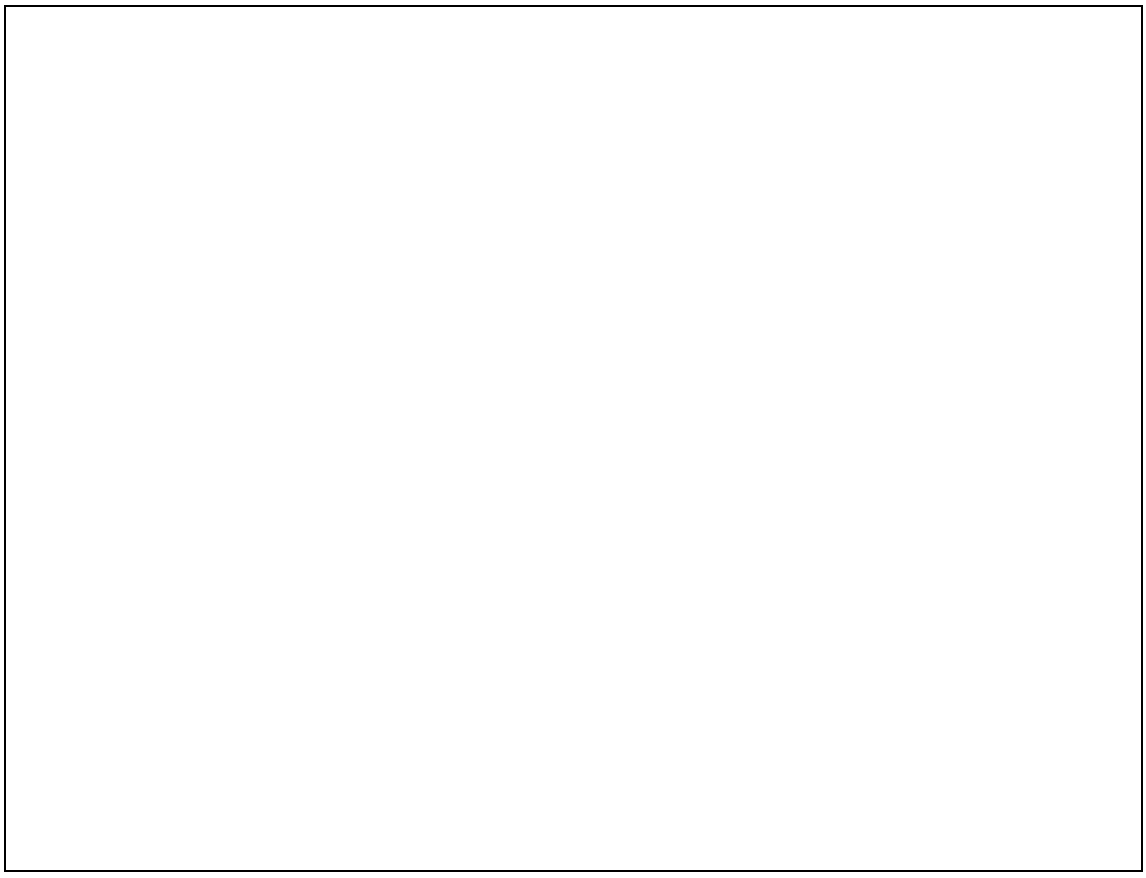
查看 runtime 的类型，发现是 object 类型，也就是字符串，所以，先进行数据转化。



5.3.1 电影时长（绘制电影时长直方图）



5.3.2 发行时间（绘制每月电影数量和单片平均票房）

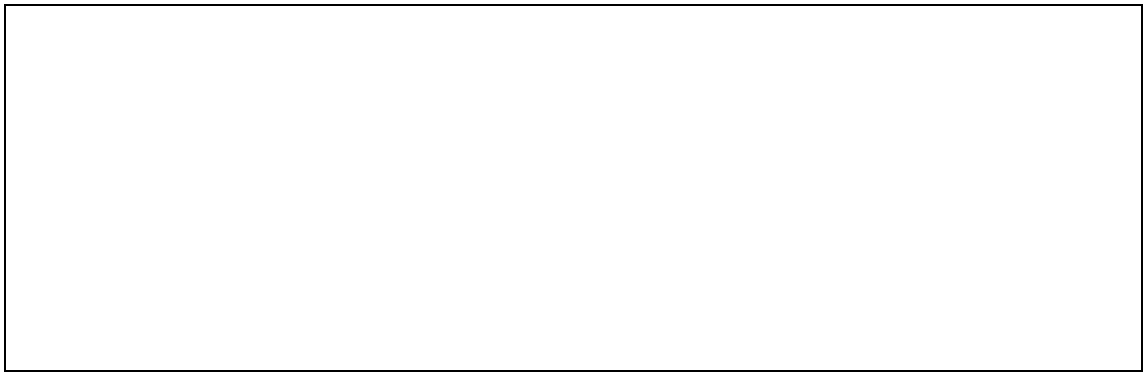


5.4 where

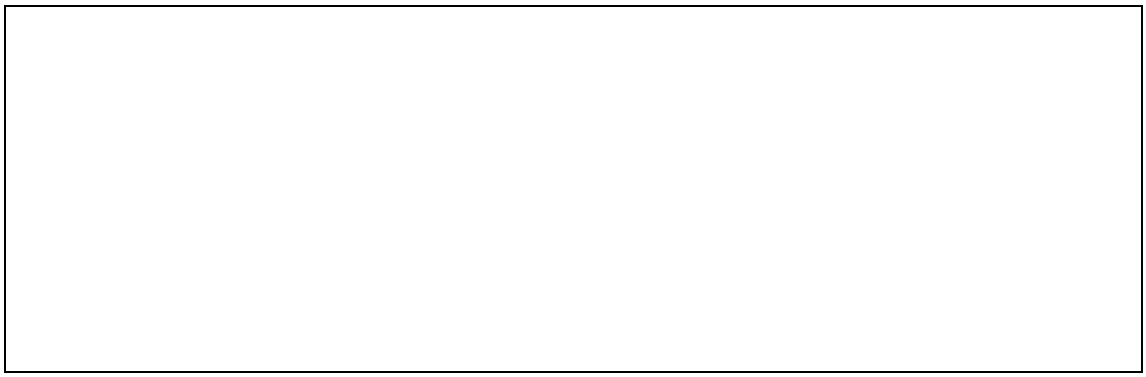
本数据集收集的是美国地区的电影数据，对于电影的制作公司以及制作国家，在本次的故事背景下不作分析。

5.5 who

5.5.1 分析票房分布及票房 Top10 的导演



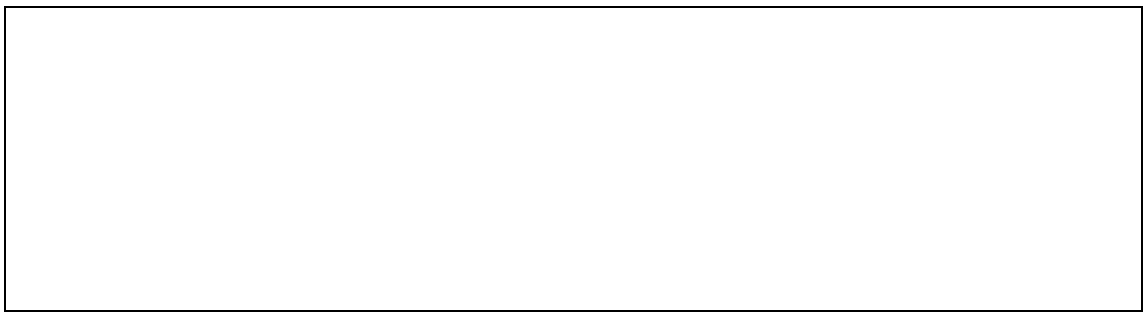
5.5.2 分析评分分布及评分 Top10 的导演



5.6 how

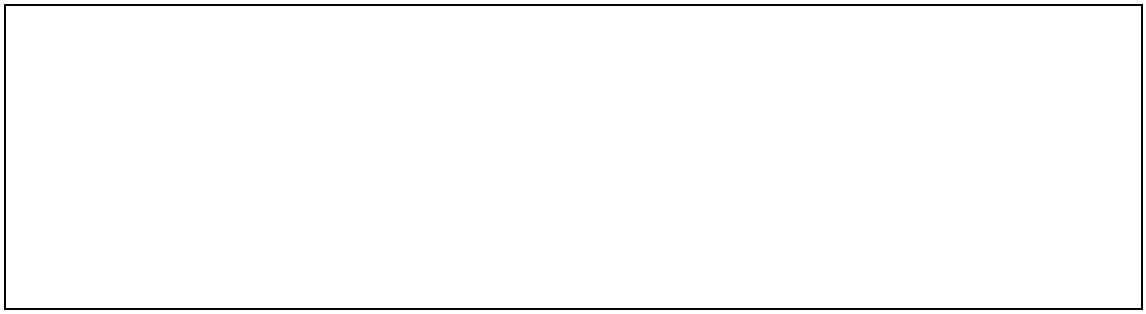
5.6.1 原创 VS 改编占比（饼图）

5.6.2 原创 VS 改编预算/利润率（组合图）

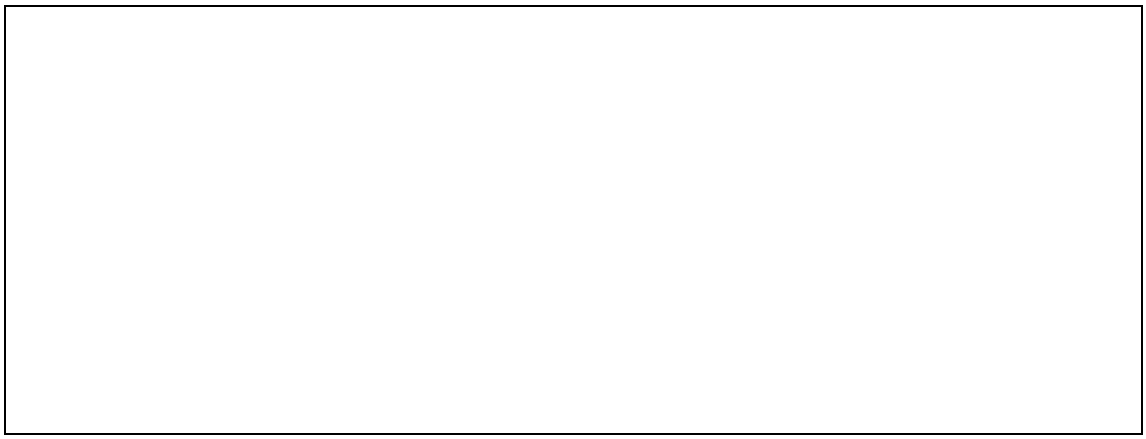


5.7 how much

5.7.1 计算相关系数（票房相关系数矩阵）



5.7.2 票房影响因素散点图



6 对附件的 tmdb_1000_predicted.csv 的预测结果