1. 目的

目的：分析电影市场情况并预测观众群对“四合影业”计划 投拍的电影《青春的竞赛》的评分

1. 准备数据

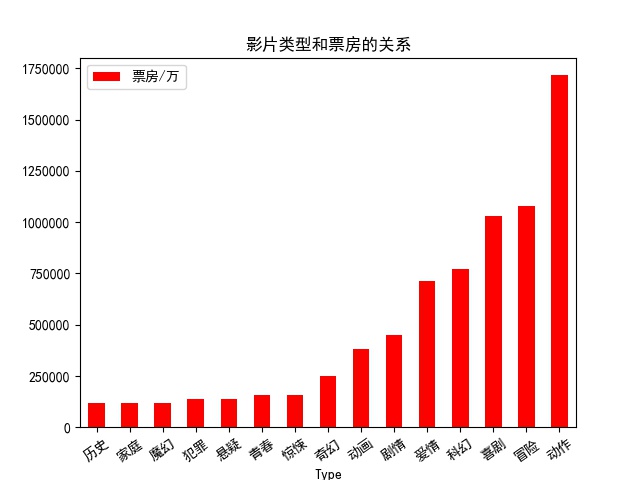
文件夹：D:\安徽赛项说明会\样题\arg\arg04

|  |  |
| --- | --- |
| 文件名称 | 提供的数据 |
| film-csv.txt | 电影名称 导演 演员 影片类型 主角 票房 评分等数据 |
| id04.txt | 五个会员 |
| score.log | 提供电影名称 会员 会员对应的评分 |

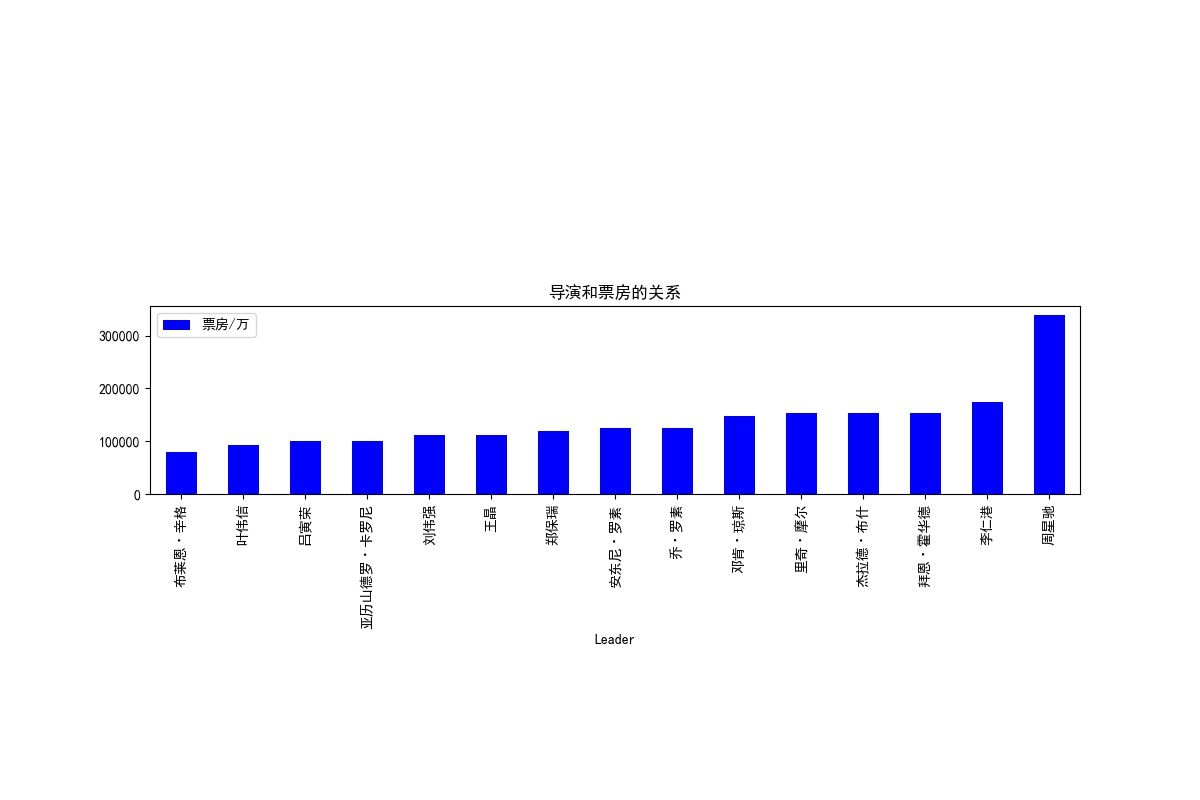
1. 数据分析
2. 去重 ：信息重复的数据
3. 去空 ：主要信息缺失为空的
4. 除去脏数据 ：与电影信息无关的信息

|  |  |
| --- | --- |
| 导包 | 主要使用的方法 |
| re | Split,sub ,search ,findall |
| numpy | Mean,median,max,mean |
| pandas | Read\_csv,query,isin,DataFrame,groupby,  Sort\_values,reset\_index,concat等 |

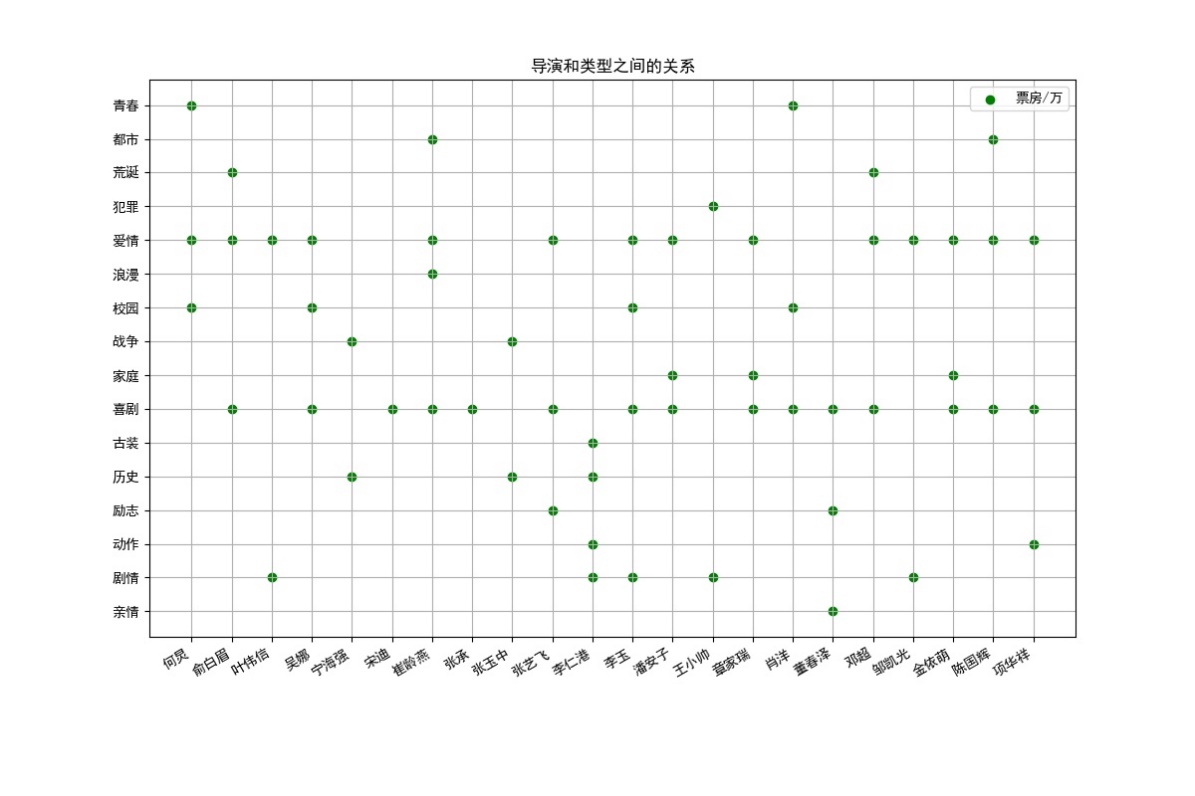
1. 绘图
2. 各种类型片票房收入比较



1. 导演票房收入比 较



1. 导演执导过的影片类型



1. 预测

预测的数据:演员:邓超，导演：周星驰，类型：喜剧

预测的结果是：投资电影的评分

分析过程：1.将得到的数据导演，类型，演员进行序列化(变成数字)，使用pandas中的query查询演员:邓超，导演：周星驰，类型：喜剧 得到对应的序列数

2.通过五个ueserid 找到对应的电影名称，并进行匹配，得到Xtrain[导演，类型，演员]，ytrain[评分]

训练模型:

regr = DecissionTreeRegression()#通过调节最大深度和修剪枝叶提高准确率

regr.fit(Xtrain,ytrain)

regr.predit([[邓超，周星驰，喜剧]])

注：分为测试集和训练集，通过调节参数提高预测的准确率

得到预测结果与大众评分如下：

|  |  |
| --- | --- |
| 大众评分 | 评分结果（单位/分） |
| 最大值 | 9.20 |
| 最小值 | 2.40 |
| 中位数 | 5.60 |
| 平均数 | 5.56 |

|  |  |
| --- | --- |
| 会员预测评分 | 评分结果（单位/分） |
| 最大值 | 9.90 |
| 最小值 | 4.60 |
| 中位数 | 8.45 |
| 平均数 | 7.44 |

总结:大众喜欢看的影片类型票房比较高的有:动作，冒险，喜剧，科幻，爱情

大众喜欢看的影片类型票房比较高的有：周星驰，李仁港，霍华德

导演们拍摄最多的电影类型有: 动作, 喜剧, 爱情

建议:可以让周星驰，尝试拍摄动作，冒险，喜剧，科幻，爱情之类的电影。

1. 本题完毕