douban-master

功能

数据获取: 使用爬虫工具, 在豆瓣TOP250榜单, 猫眼网票房排行榜上爬取电影相关数据, 如评分,票房等

数据持久化: 使用pandas中的DataFrame存储csv的方式和MySQL关系型数据库存储两种方式分别实现持久化

可视化分析: 从持久化的数据中选取相应数据的关系进行可视化分析

票房预测: 通过可视化分析得到的结论, 选取可能影响票房的因素, 建立预测模型和算法, 进行预测

文件结构

文件	描述
main.py	数据爬虫及持久化的主函数
movie_basic.py	豆瓣TOP250列表页爬取
movie_detail.py	豆瓣电影详情内页爬取
database.py	数据库连接操作及查询接口
attachfile.py	静态内容，如请求头headers等
visualization_sql.ipynb	数据可视化，数据使用SQL查询方式
visualization_pandas.ipynb	数据可视化，数据使用pandas聚合等方式
predict.ipynb	票房预测模型的建立和预测举例
/html	存放爬取的html文件
/csv	存放持久化pandas处理的dataframe数据
/result	存放可视化结果，及数据库内容截图等

技术栈

Python爬虫与数据处理: requests, lxml, re, pandas

数据持久化: pymysql, pandas, MySQL

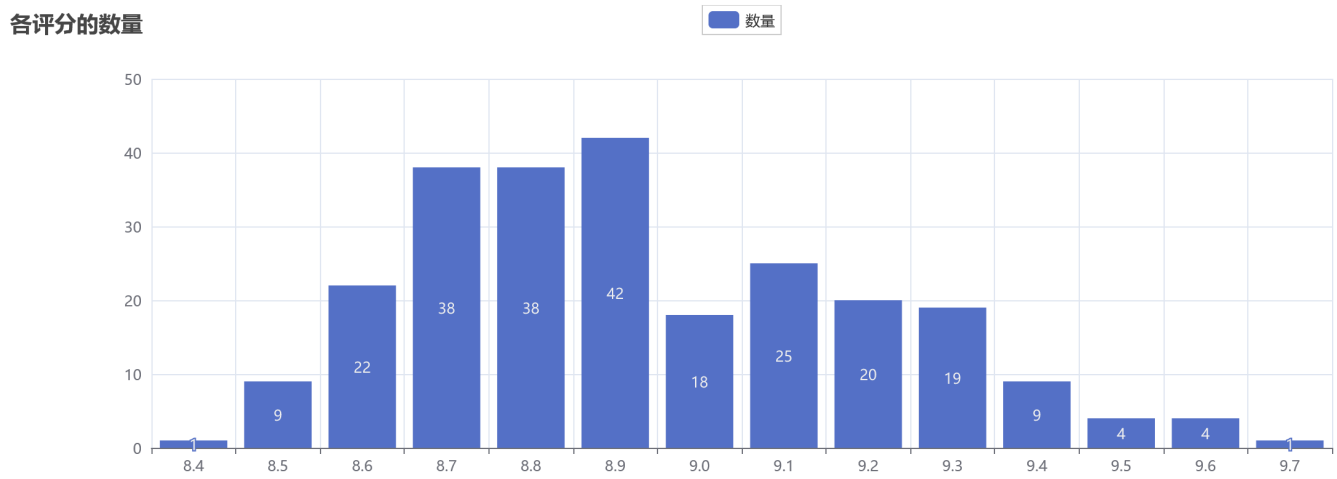
数据清洗: pandas, MySQL (实际上没做)

可视化分析: pyecharts, matplotlib, SQL, pandas

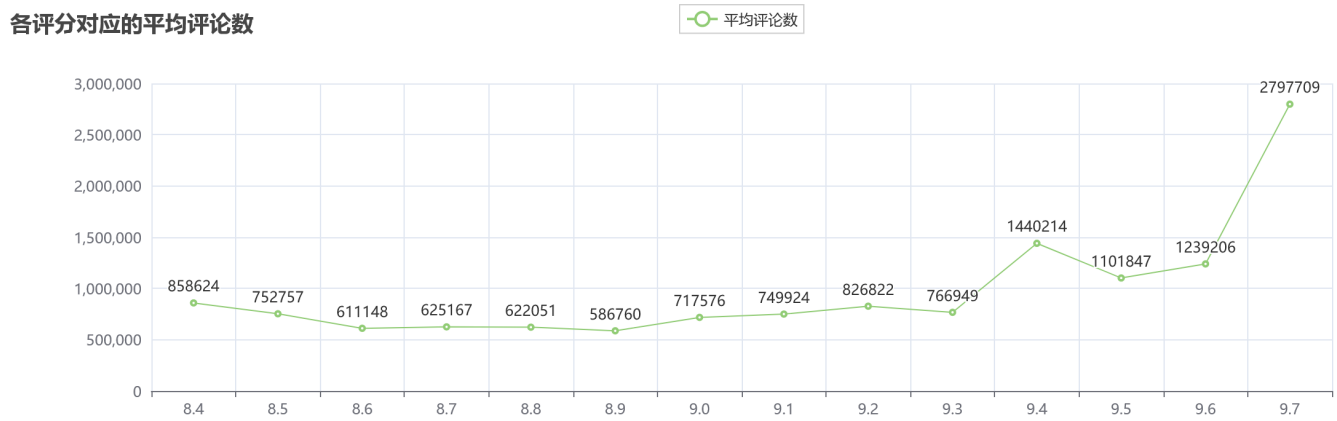
模型预测: sklearn, numpy, matplotlib

可视化举例

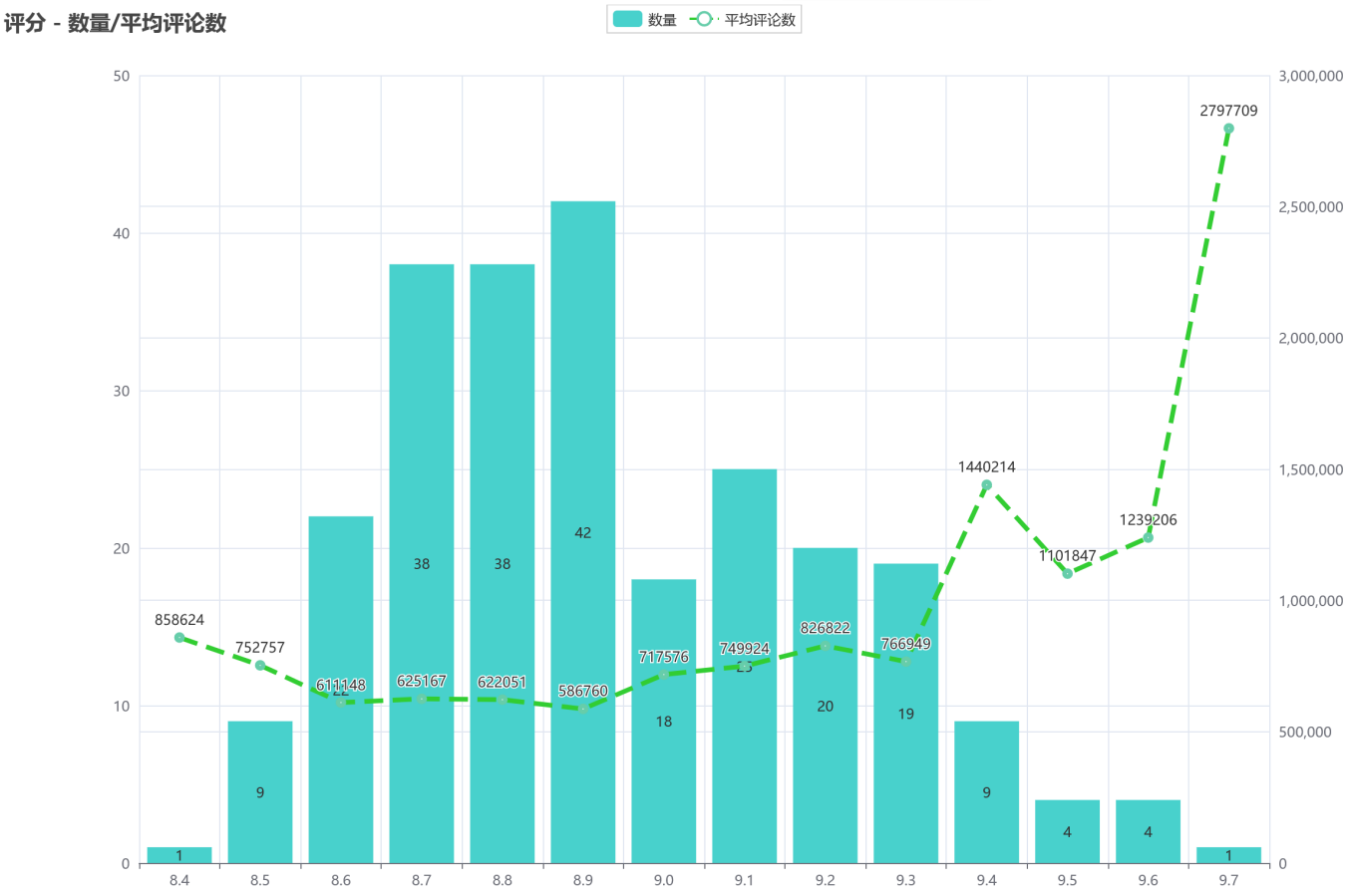
各评分的数量



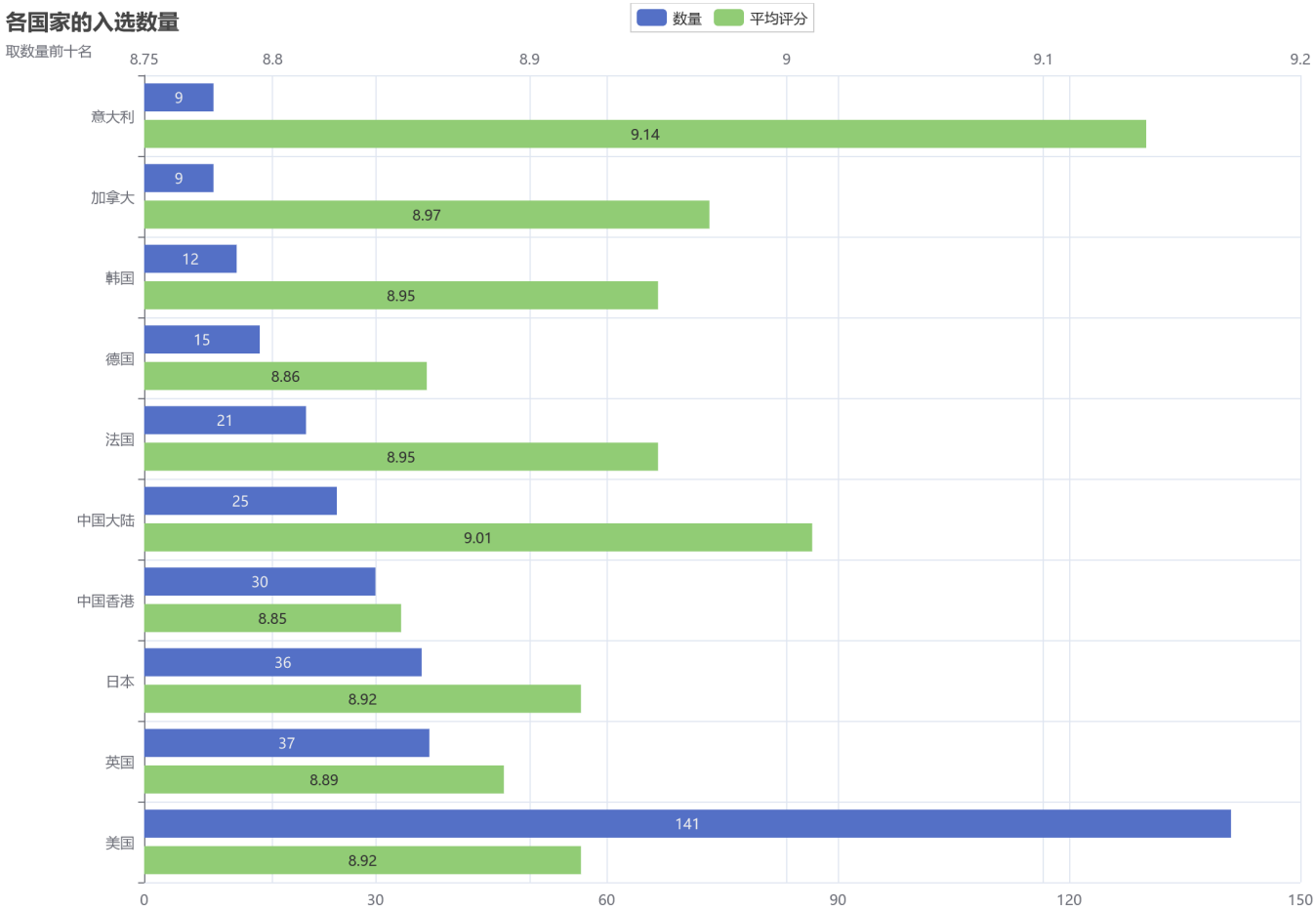
各评分对应的平均评论数



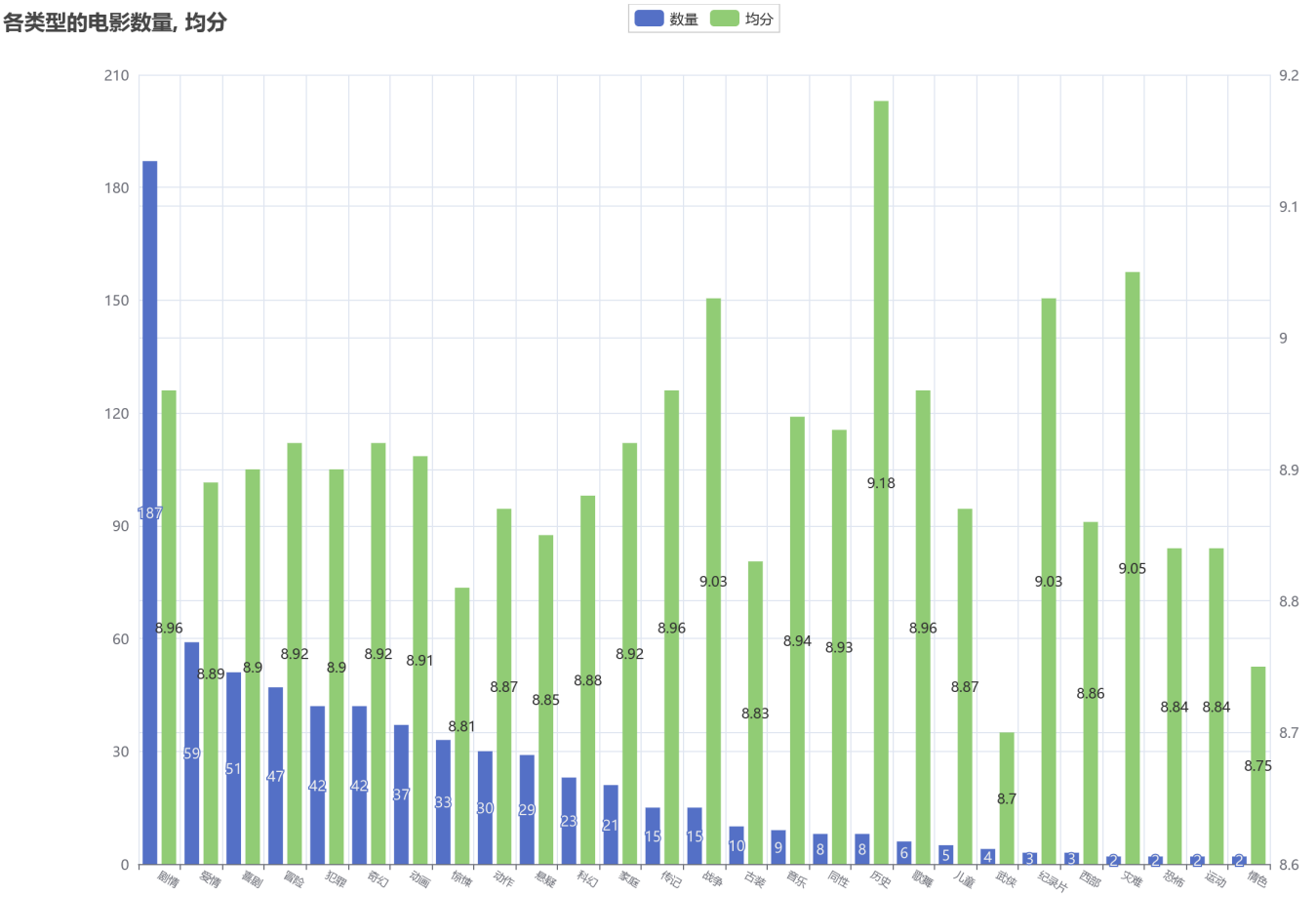
评分 - 数量/平均评论数



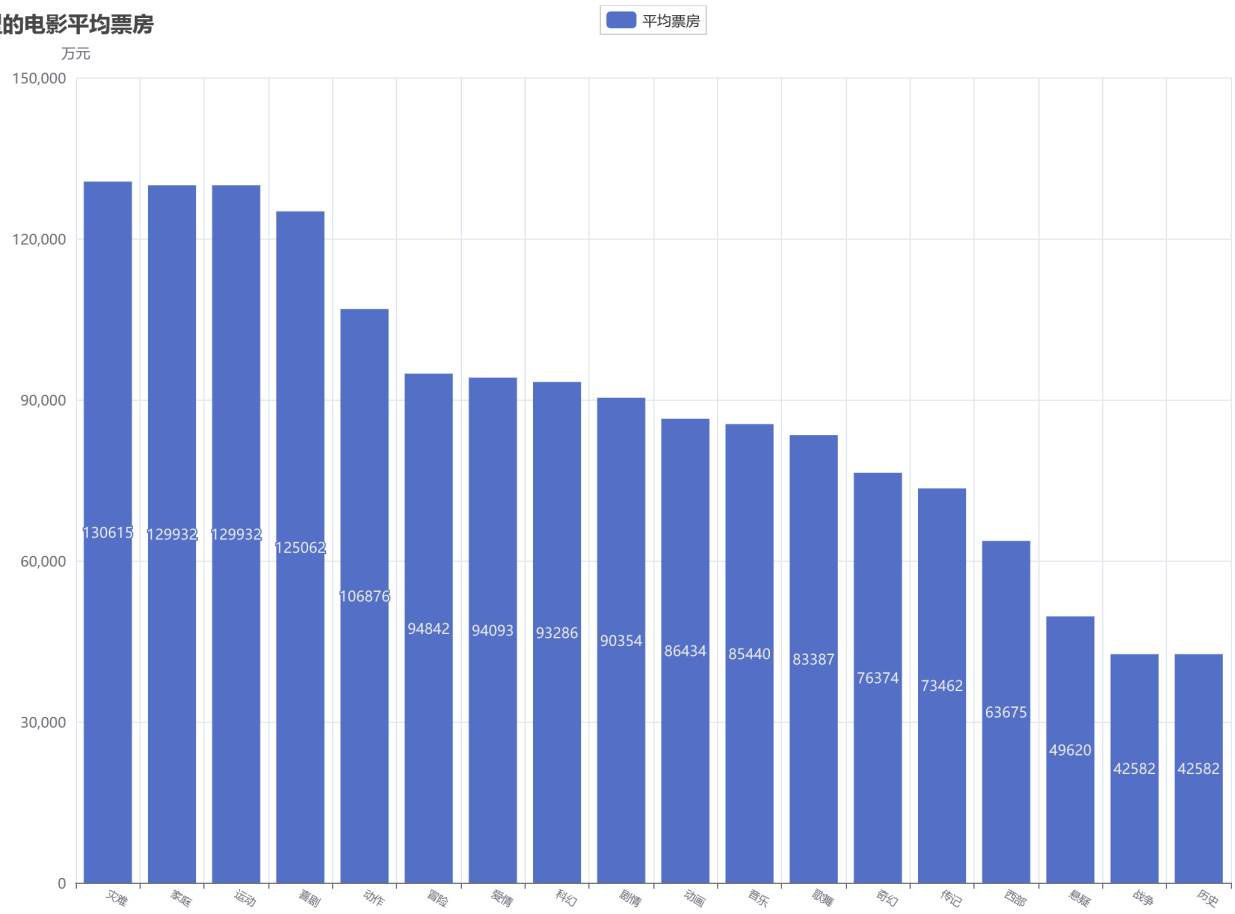
各国家的入选数量



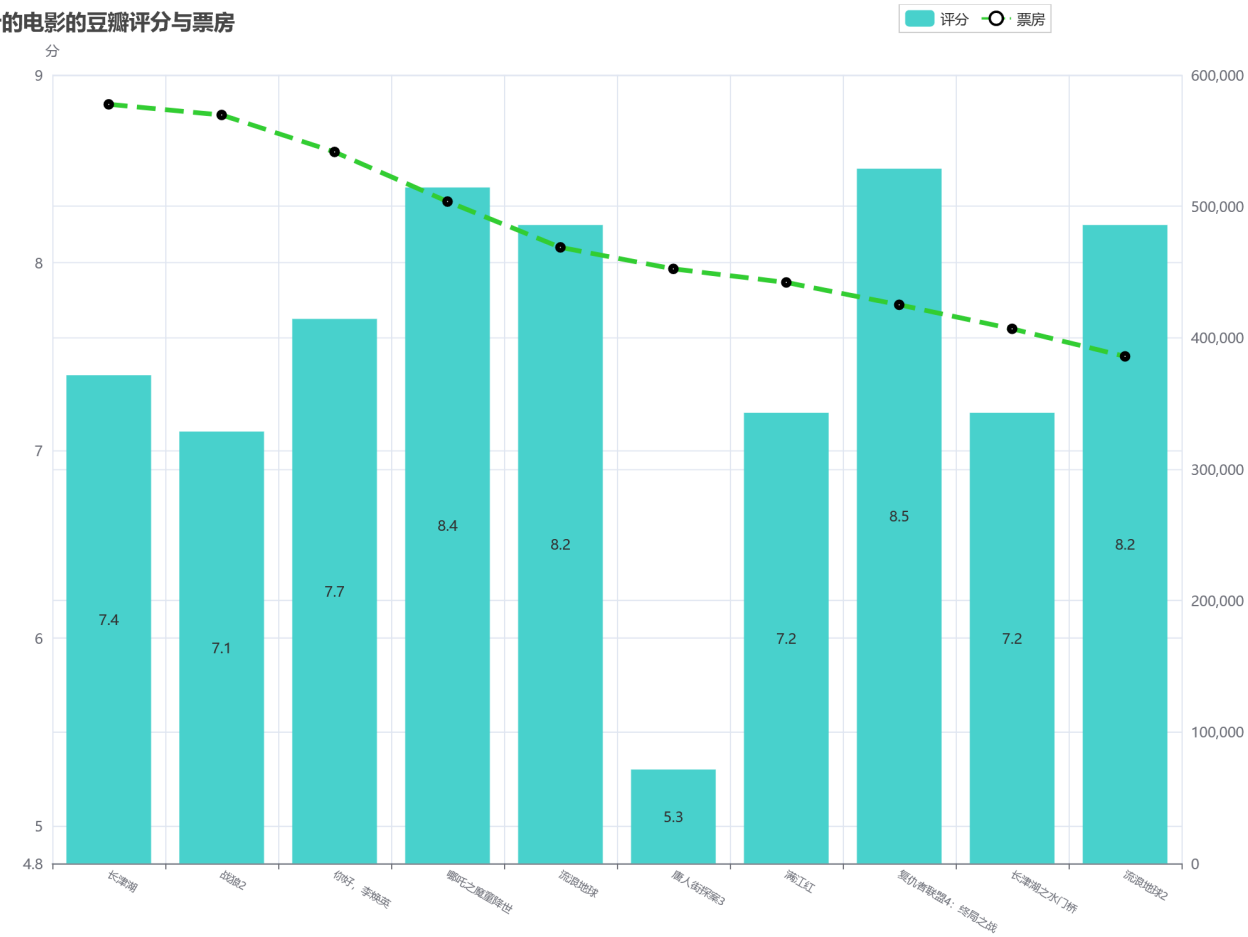
各类型的电影数量, 均分

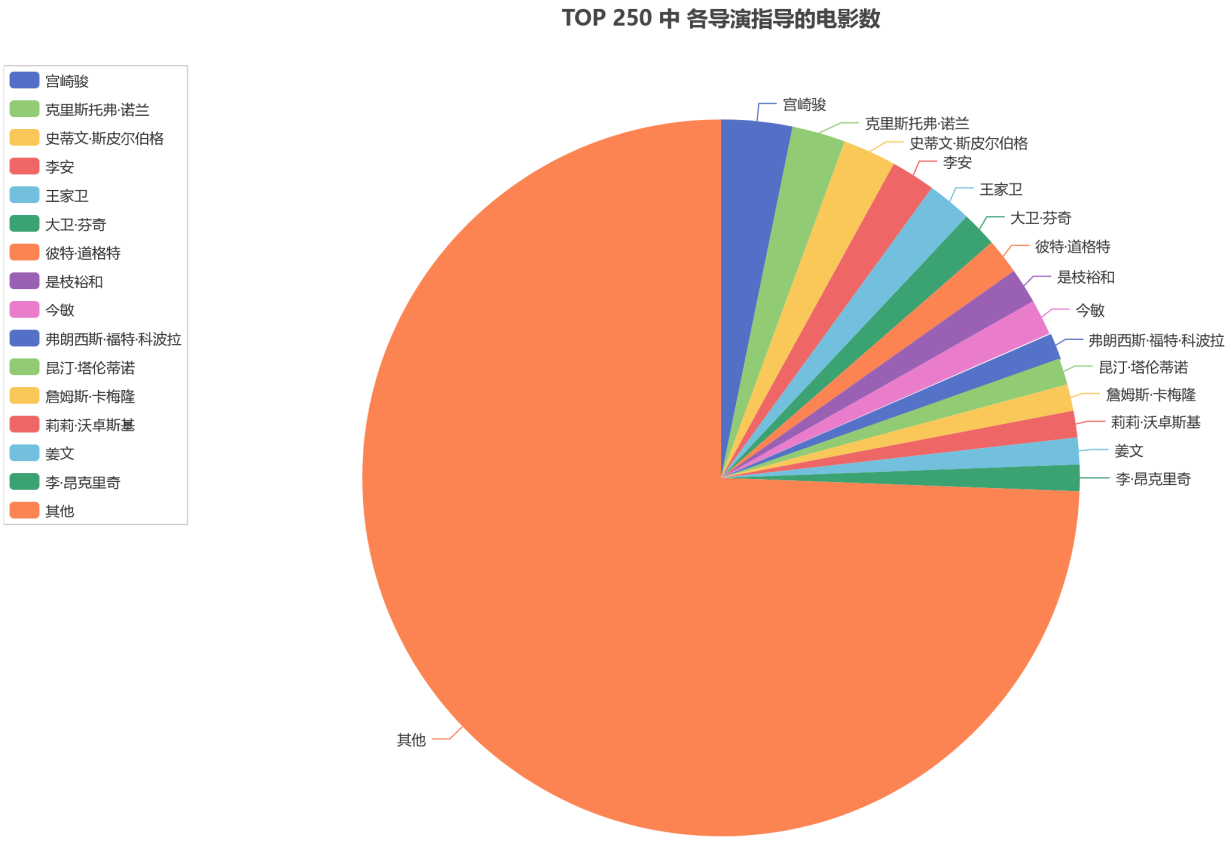


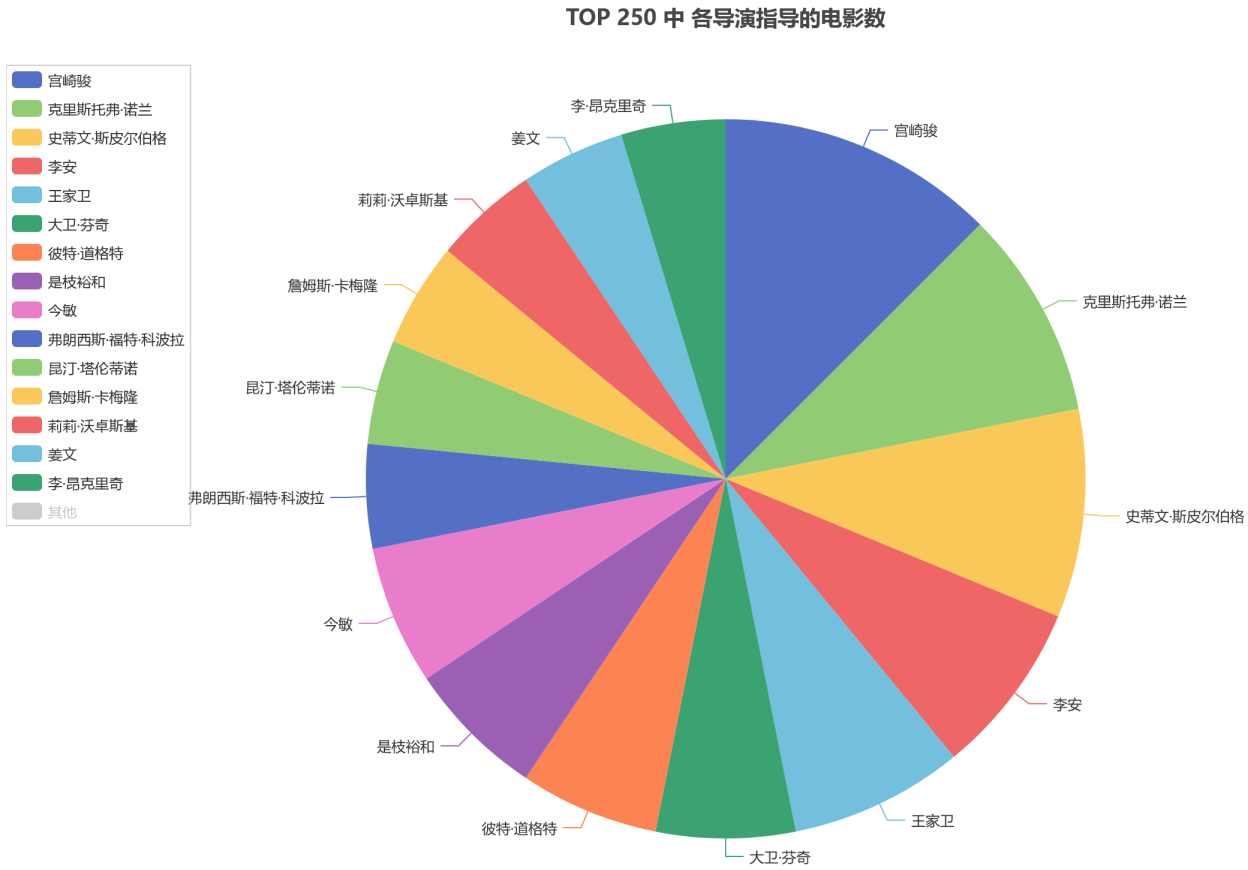
各类型的电影平均票房

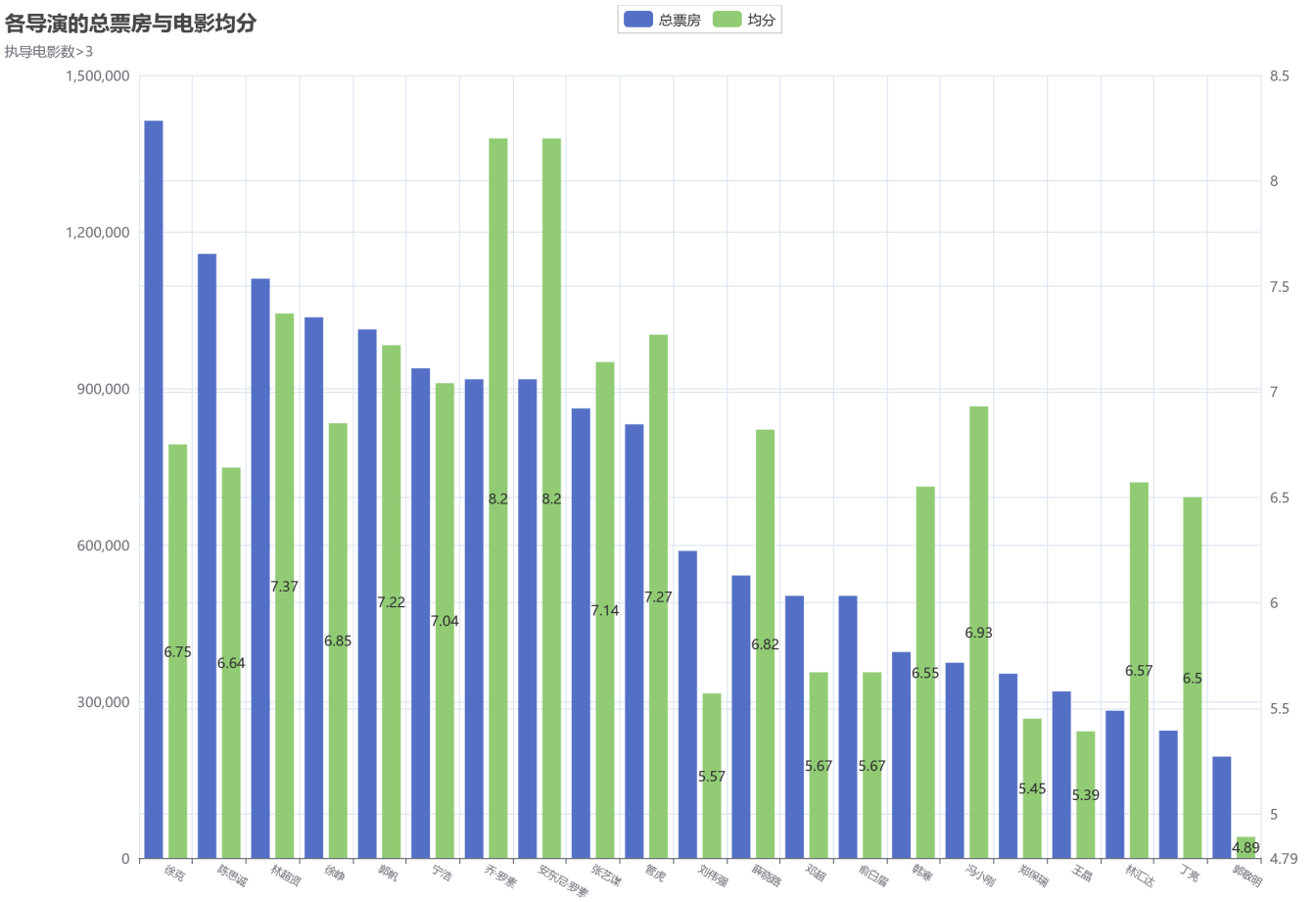


票房前十的电影的豆瓣评分与票房

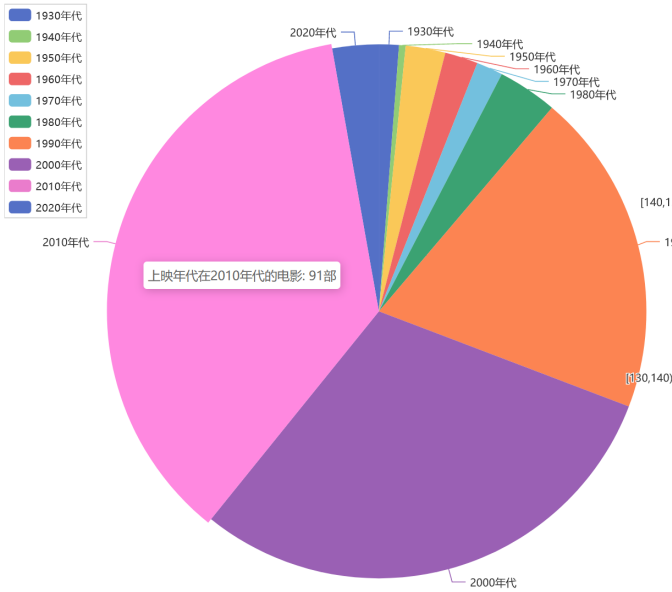



TOP 250 中 各导演指导的电影数
宫崎骏
克里斯托弗·诺兰
史蒂文·斯皮尔伯格
李安
王家卫
大卫·芬奇
彼特·道格特
是枝裕和
今敏
弗朗西斯·福特·科波拉
昆汀·塔伦蒂诺
詹姆斯·卡梅隆
莉莉·沃卓斯基
姜文
李·昂克里奇
其他

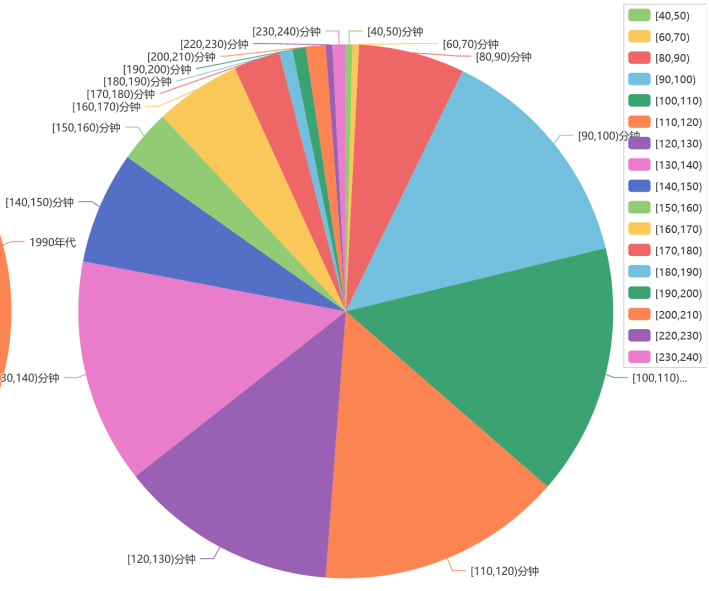

TOP 250 中 各导演指导的电影数
宫崎骏
克里斯托弗·诺兰
史蒂文·斯皮尔伯格
李安
王家卫
大卫·芬奇
彼特·道格特
是枝裕和
今敏
弗朗西斯·福特·科波拉
昆汀·塔伦蒂诺
詹姆斯·卡梅隆
莉莉·沃卓斯基
姜文
李·昂克里奇
其他

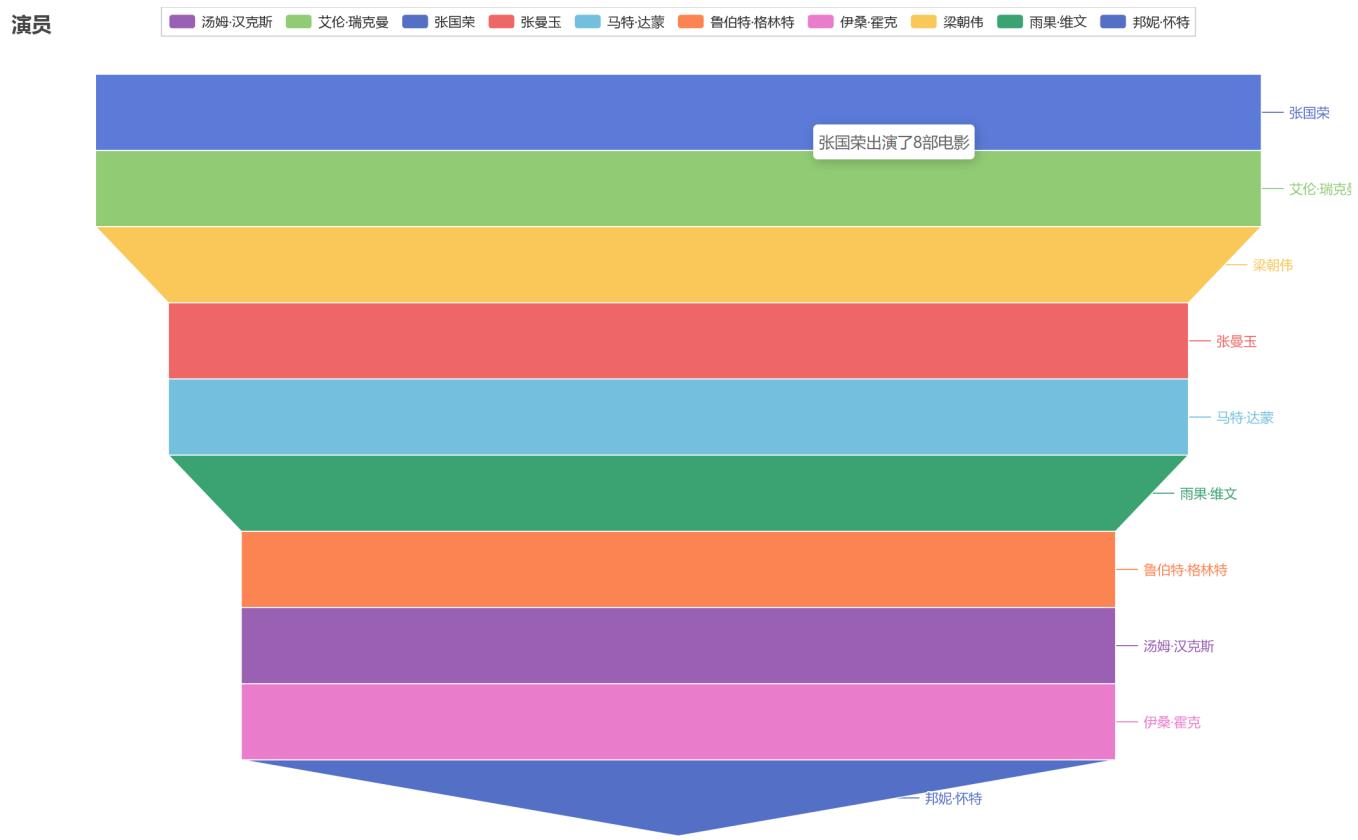


不同年代上映的电影数量



不同时长的电影数量





票房预测举例

单位/万元

```
movie = [
    ['剧情', '历史'],
    ['吴京', '包贝尔', '易烊千玺', '邓超', '欧豪', '雷佳音', '郭京飞'],
    ['徐克', '吴京'],
    2
]

predict(movie)
```

[17] Python

```
(4.6808694950739635, 406733, 2022, '长津湖之水门桥')
(3.846339605633867, 577534, 2021, '长津湖')
(3.616970526244077, 165207, 2017, '西游伏妖篇')
(3.254676920420864, 54469, 2015, '战狼')
(3.2450937914128564, 88348, 2014, '智取威虎山')

306848.3916073195
```

```
movie = [
    ['爱情', '喜剧'],
    ['韩庚', '郑恺', '于文文', '刘雅瑟', '刘天爱'],
    ['毋羽斗'],
    10
]

predict(movie)
```

[18] Python

```
(8.941589026166287, 194190, 2017, '前任3：再见前任')
(3.0211400439542415, 58861, 2014, '匆匆那年')
(2.668806447200494, 71902, 2013, '致我们终将逝去的青春')

169230.68515821127
```