《Python语言数据分析基础》课程期末大作业

——对豆瓣电影TOP250数据分析

18电气专升本 洪浩特

## 一、为什么做数据分析？

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。这一过程也是质量管理体系的支持过程。在实用中，数据分析可帮助人们作出判断，以便采取适当行动。 数据分析的数学基础在20世纪早期就已确立，但直到计算机的出现才使得实际操作成为可能，并使得数据分析得以推广。数据分析是数学与计算机科学相结合的产物。

## 为什么做豆瓣电影数据分析？

豆瓣网站是国内较大的评论（[书评](https://baike.baidu.com/item/%E4%B9%A6%E8%AF%84/746506" \t "https://baike.baidu.com/item/%E8%B1%86%E7%93%A3%E7%BD%91/_blank)、[影评](https://baike.baidu.com/item/%E5%BD%B1%E8%AF%84/10364061" \t "https://baike.baidu.com/item/%E8%B1%86%E7%93%A3%E7%BD%91/_blank)、[乐评](https://baike.baidu.com/item/%E4%B9%90%E8%AF%84" \t "https://baike.baidu.com/item/%E8%B1%86%E7%93%A3%E7%BD%91/_blank)）网站致，其力于帮助都市人群发现生活中有用的事物。使用豆瓣的人，会发现这里能找到许多冷僻的书、电影、音乐，排在豆瓣排行榜上的大多是与市面畅销流行榜不一样的陌生面孔。对豆瓣电影数据能很好的了解当今最流行，最好看的电影是什么。根据豆瓣所有的电影，分析各国各地区各类别时间年份评分数量等各个参数之间的联系，大体上进行分析。

## 豆瓣电影数据分析应用

随着电影数量的越来越多的趋势，用户常常找不到自己期望的电影。使用数据分析可以帮助我们理解用户的需求以及市场的动向，快速帮助用户寻找到最符合心理期望的电影，节省用户时间。其次可以根据上榜电影的类型帮助制作人制作出符合用户期望的电影。

## 四、豆瓣电影TOP250数据分析

### 1.数据来源

数据爬取自豆瓣电影网站（https://movie.douban.com）。采用Python爬虫对网站数据的采集以及过滤。

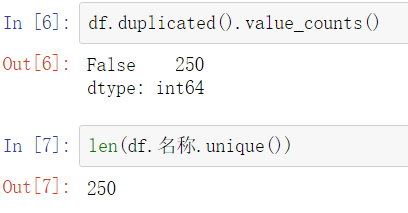
### 数据展示及数据分析结果

#### 2.1爬取结果



图 1数据部分展示

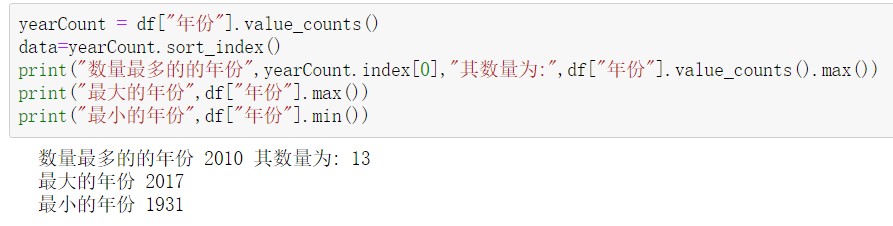
#### 2.2数据清洗



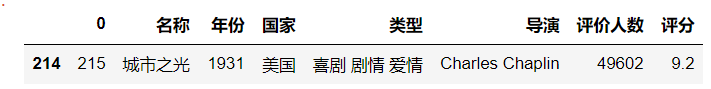
可以看到数据删除重复值后，数量没有变化说明数据没有重复值。

#### 2.3电影上映年份分析

首先分析上榜的电影年份



最老上榜的的电影是1931年的它是哪部电影呢？

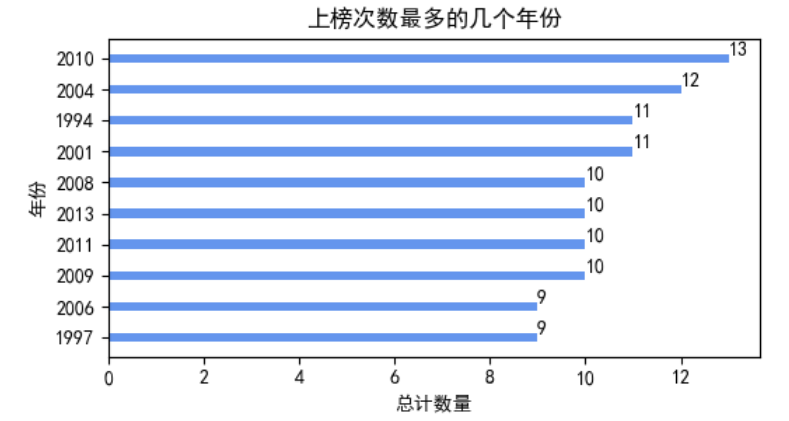


分析上榜电影数量随年份数量的变化值，如下图所示。

#### 

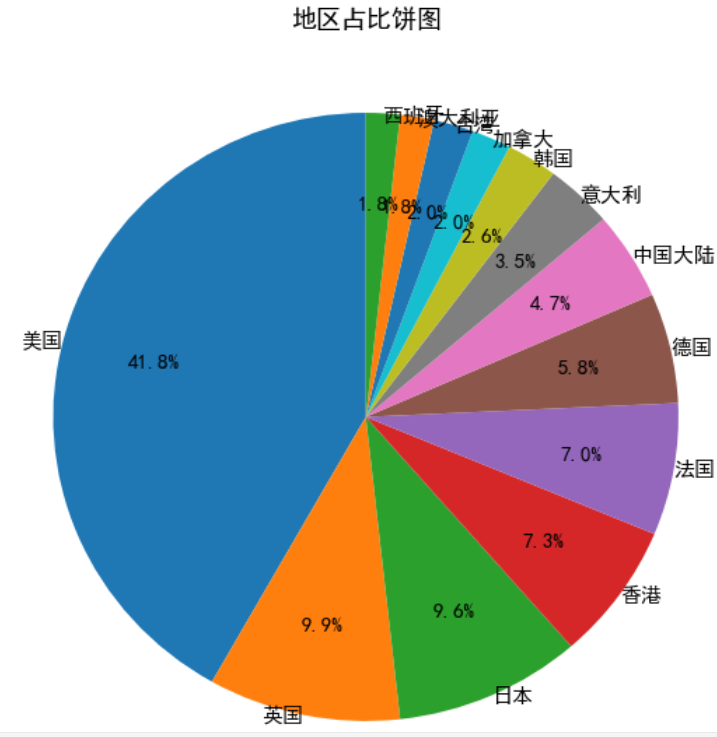
可以看到从1994年到2016年电影上榜的电影数量最多。

那么最多的数量最多的是哪几年呢？



#### 2.4电影地区分析

分析每个地区制作的电影占比，数量较少的用其他代替。



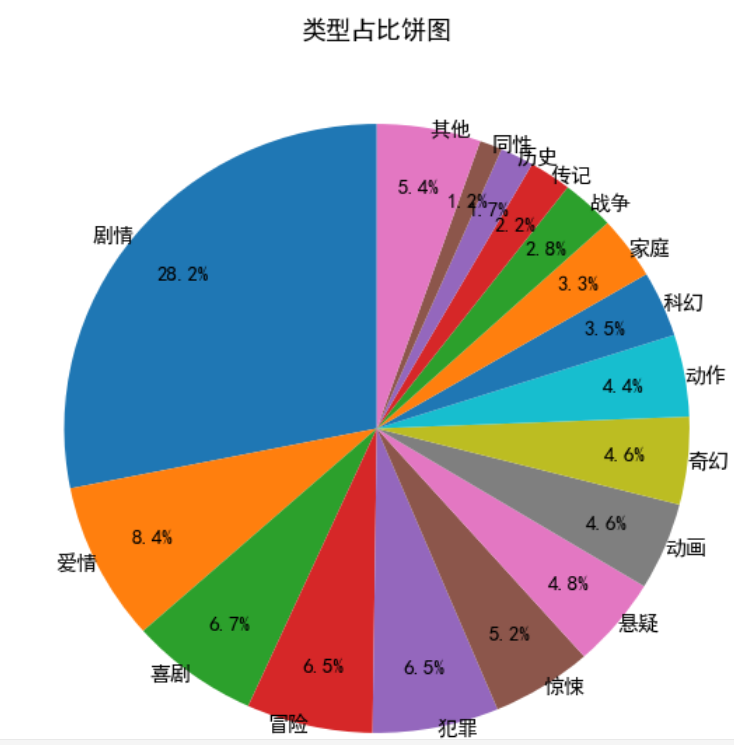
上图看到美国制作的电影数量最多远远超过其他国家的电影的电影

各个地区最高的评分是哪些电影?



#### 2.5电影类型分析

最受欢迎的类型是什么?

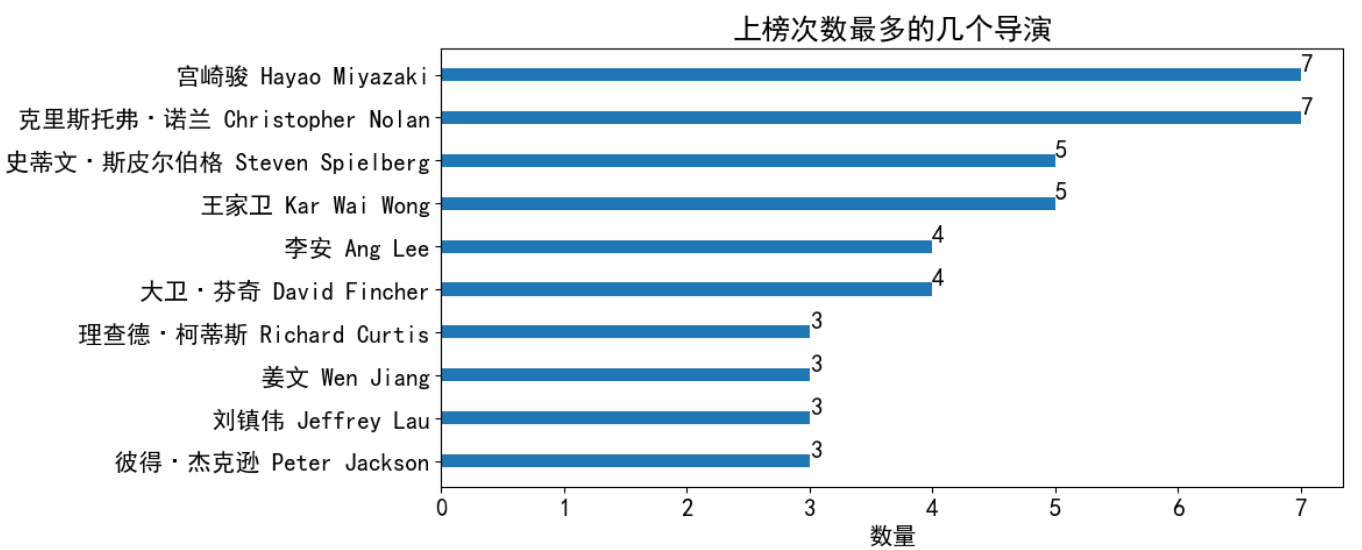


剧情片占比最多。

各种类型评分最高的电影是什么？



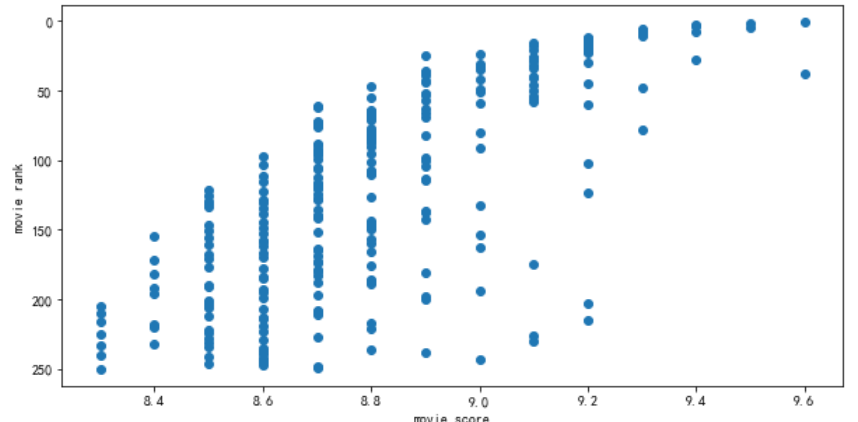
#### 2.6电影导演分析



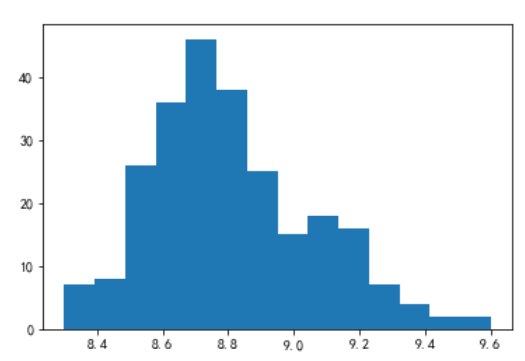
这些导演中评分最高的电影有哪些？



#### 2.7 评分和排名的关系可视化---电影排名和评分的相关性检测

 评分高的电影的排行也高，说明豆瓣电影排行很大依据是根据用户的评分而评定的机制。

#### 2.8 电影评分趋势



评分集中在8到9分之间。

#### 2.9电影推荐

9分以上并且评论超过60万的不看就浪费生命系列电影。



## 五、总结

通过对豆瓣Top250电影分析，了解很多电影知识并学会数据的清洗规范-数据存储-数据可视化-数据分析，虽然非常浅显而没有太多价值，但这也是我这种菜鸡必须需要经历的一步，完完整整，虽有各种波折，所幸全部解决，从中也学到很多，以后编代码思考也会成熟。