豆瓣电影数据分析报告

互联网数据分析体系课（案例2）

1. **分析范围**

分析对象：豆瓣电影数据，不包含电视剧、综艺等其他类型

分析范围：电影发行量、评分表现、受关注程度

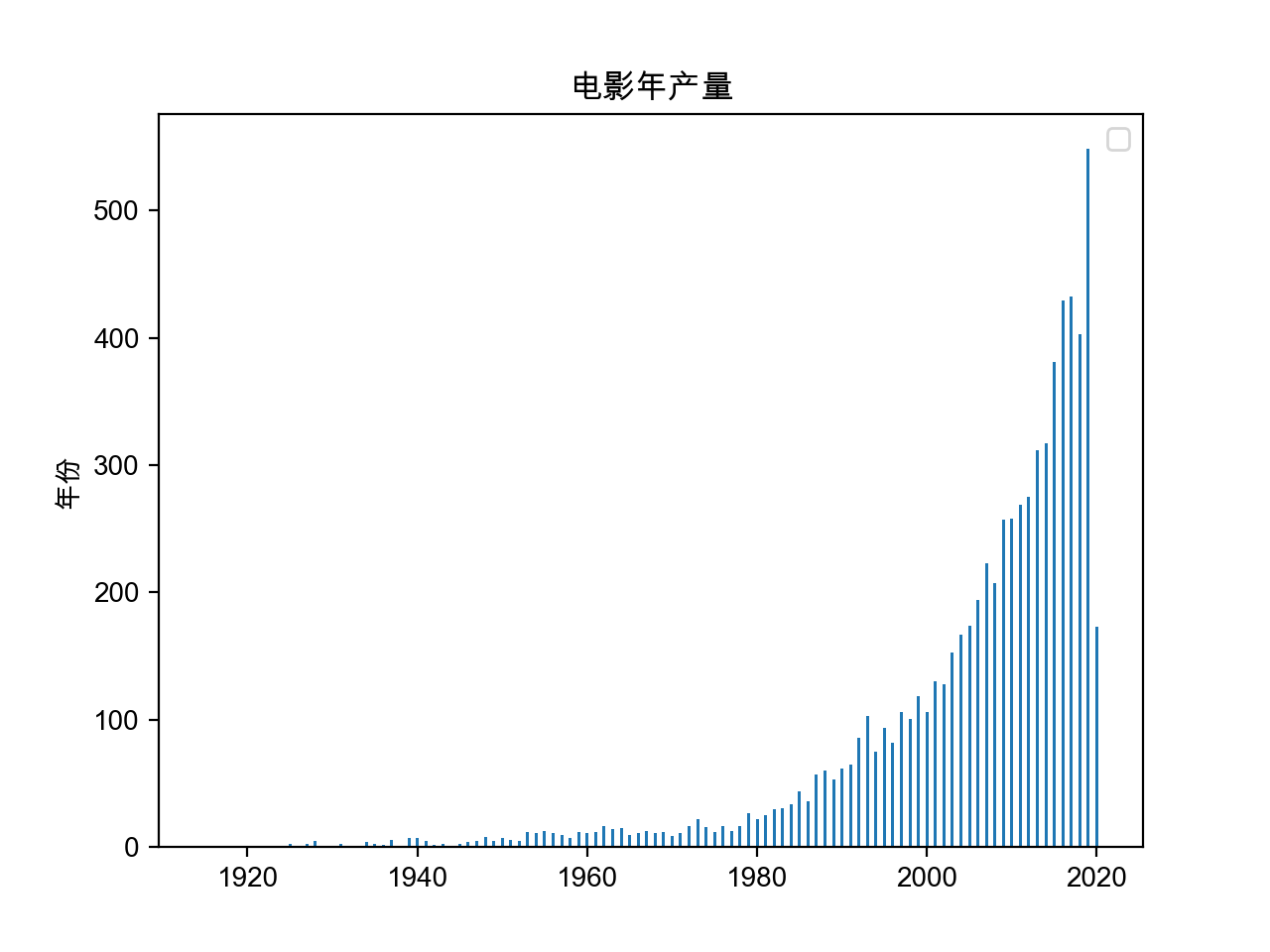
1. **数据说明**

数据来源：[www.douban.com](http://www.douban.com)，通过Python爬虫获取

数据总量：8869条电影数据

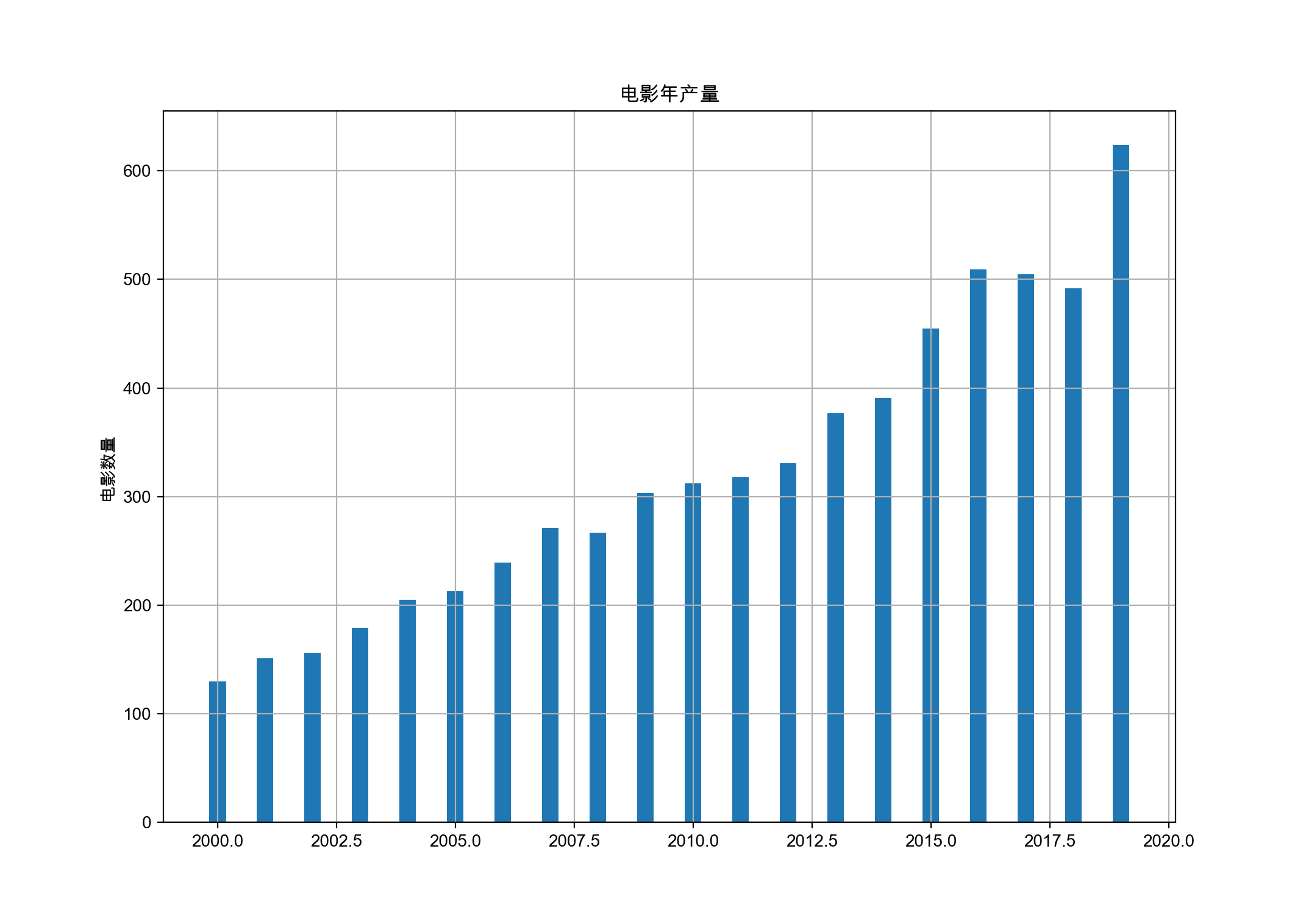
统计时间：2020年8月

1. **每年电影数量变化**



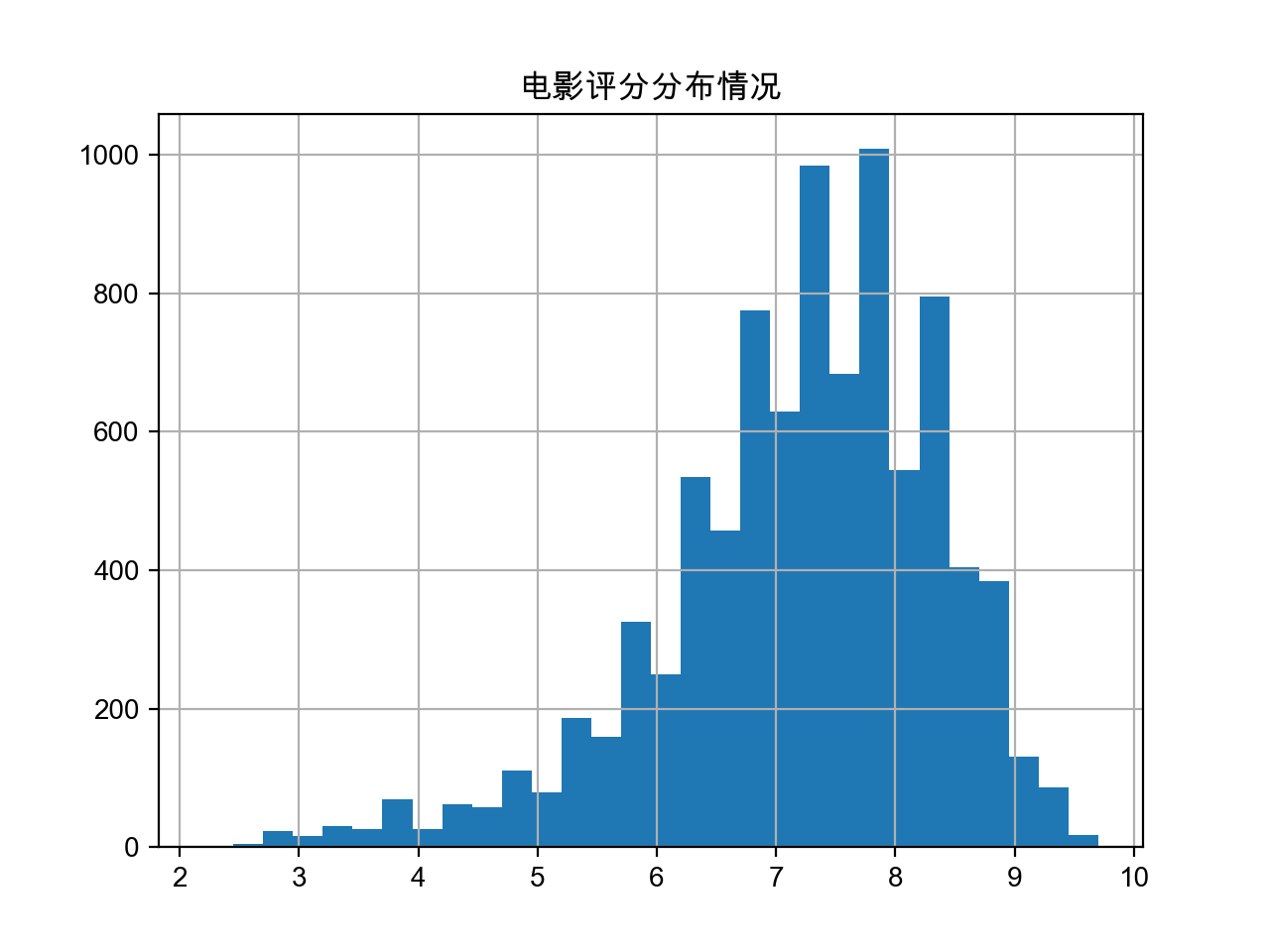
从数据来看，从1980年开始，电影年产量开始迅速增长。数据中，发行最早的是《一个国家的诞生 The Birth of a Nation》，1915年2月8日于美国上映。

我们将时间线缩短至近20年，再进一步观察。

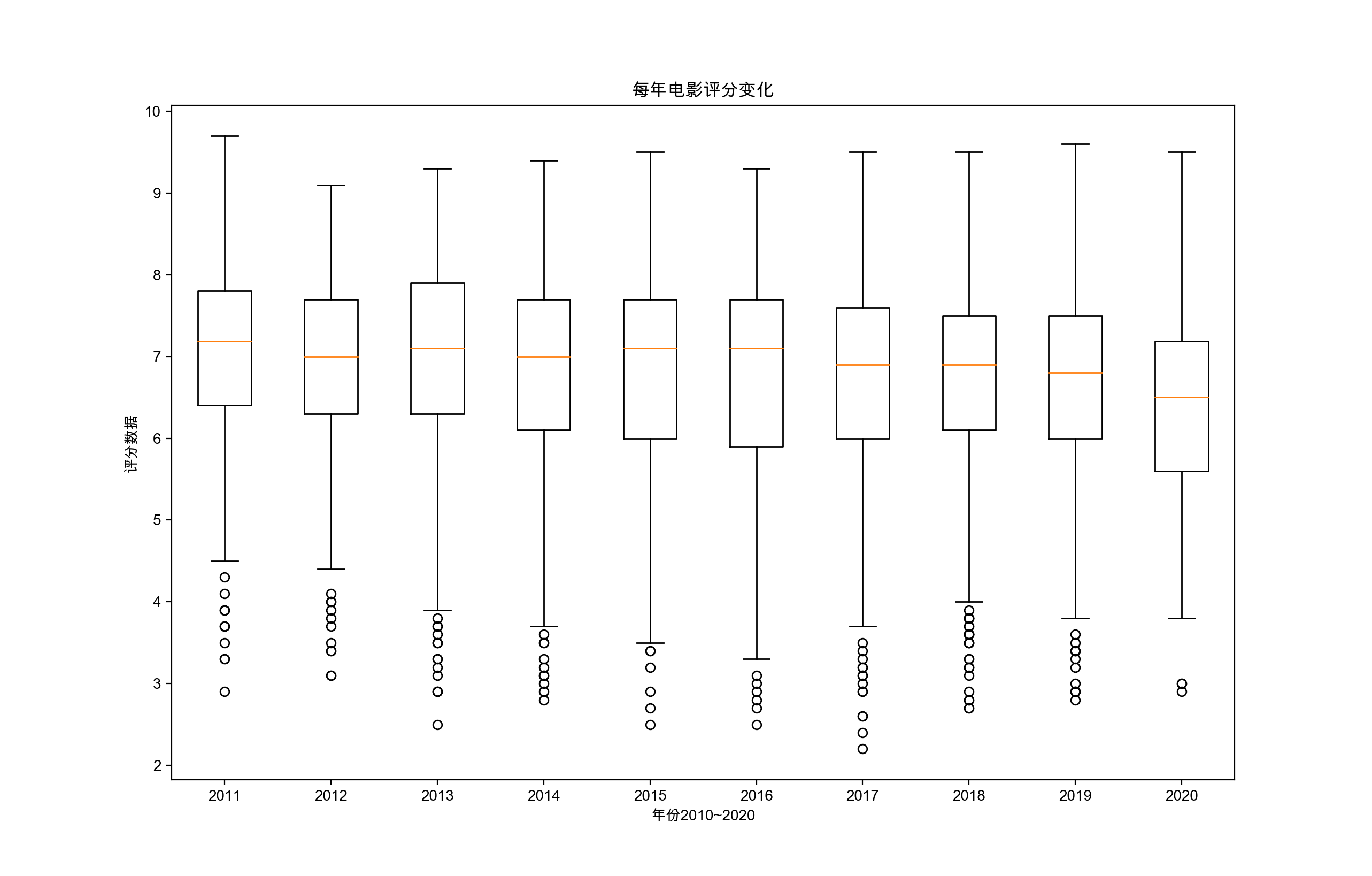


可以看到，在2019年电影发行量达到了一个峰值，全年共有624部电影上映，同比增长26.8%，市场非常火热。

1. **评分分布与变化情况**

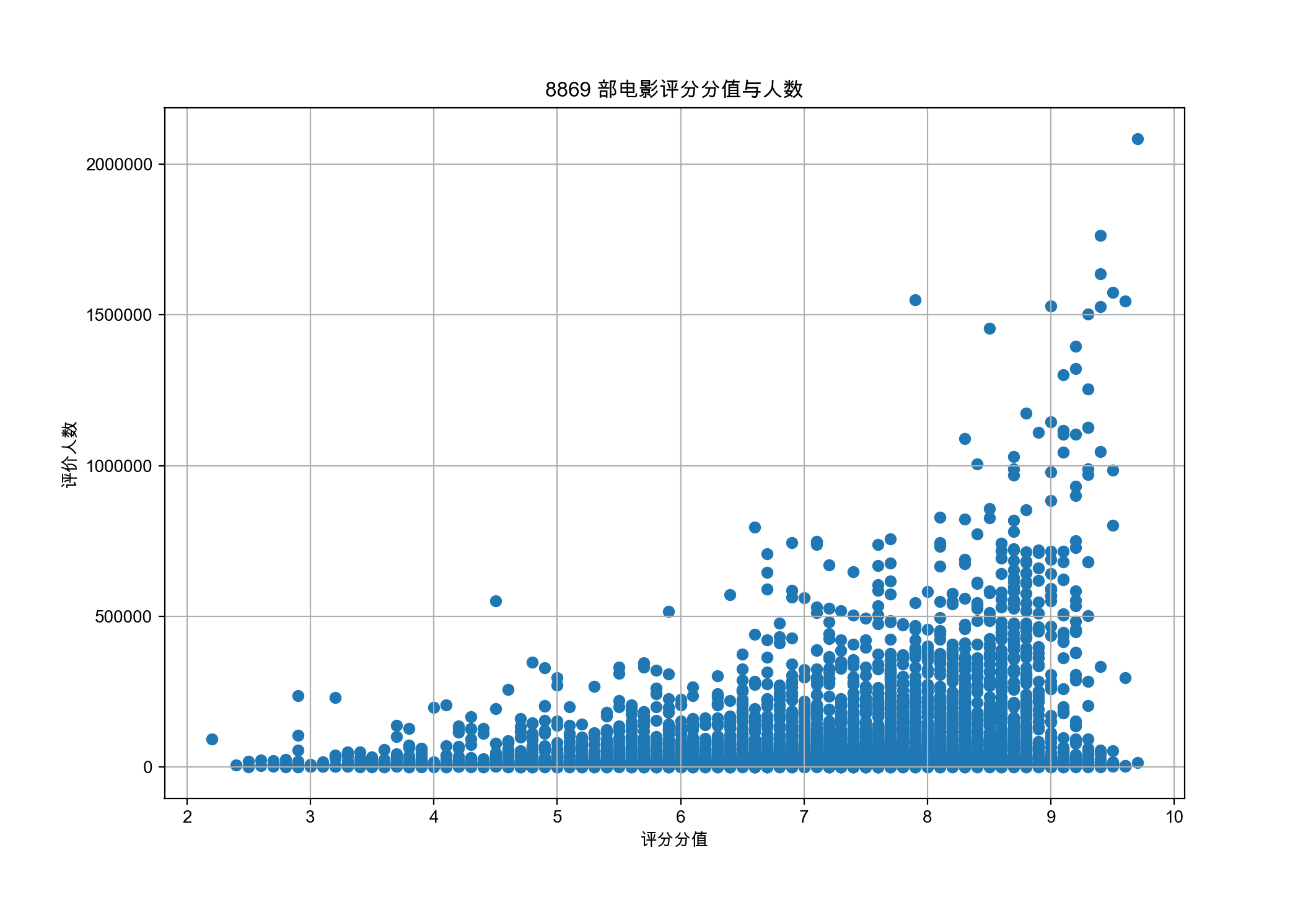
****

评分情况呈左偏分布，形态基本符合正态分布，中间区域的评分情况占多数，极端好评和差评的电影均为少数。经计算得出，电影评分均值为7.18，高于中位数6。以7.18为分界，高于这个评分的，可认为是质量不错的电影。那么，近10年的电影评分情况有什么变化吗？



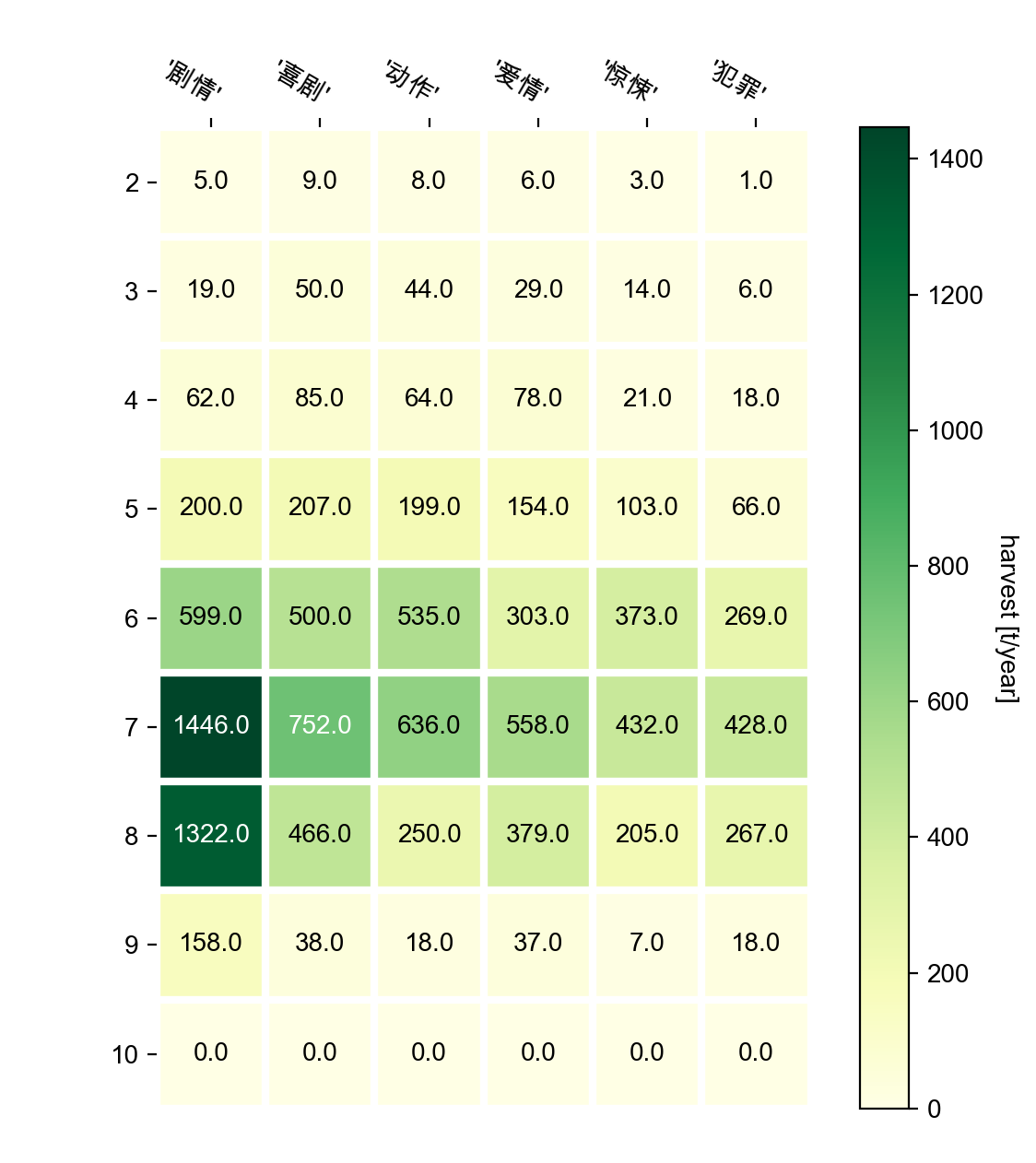
以上是每年电影评分的箱线图，橘色的线代表中位数。我们可以明显看到，随着年份的变化，橘色线所在的位置也在逐渐下移。也就意味着评分正在逐年下降。这样就意味着，电影质量越来越差，市场鱼龙混杂。

1. **评分与观看人数**



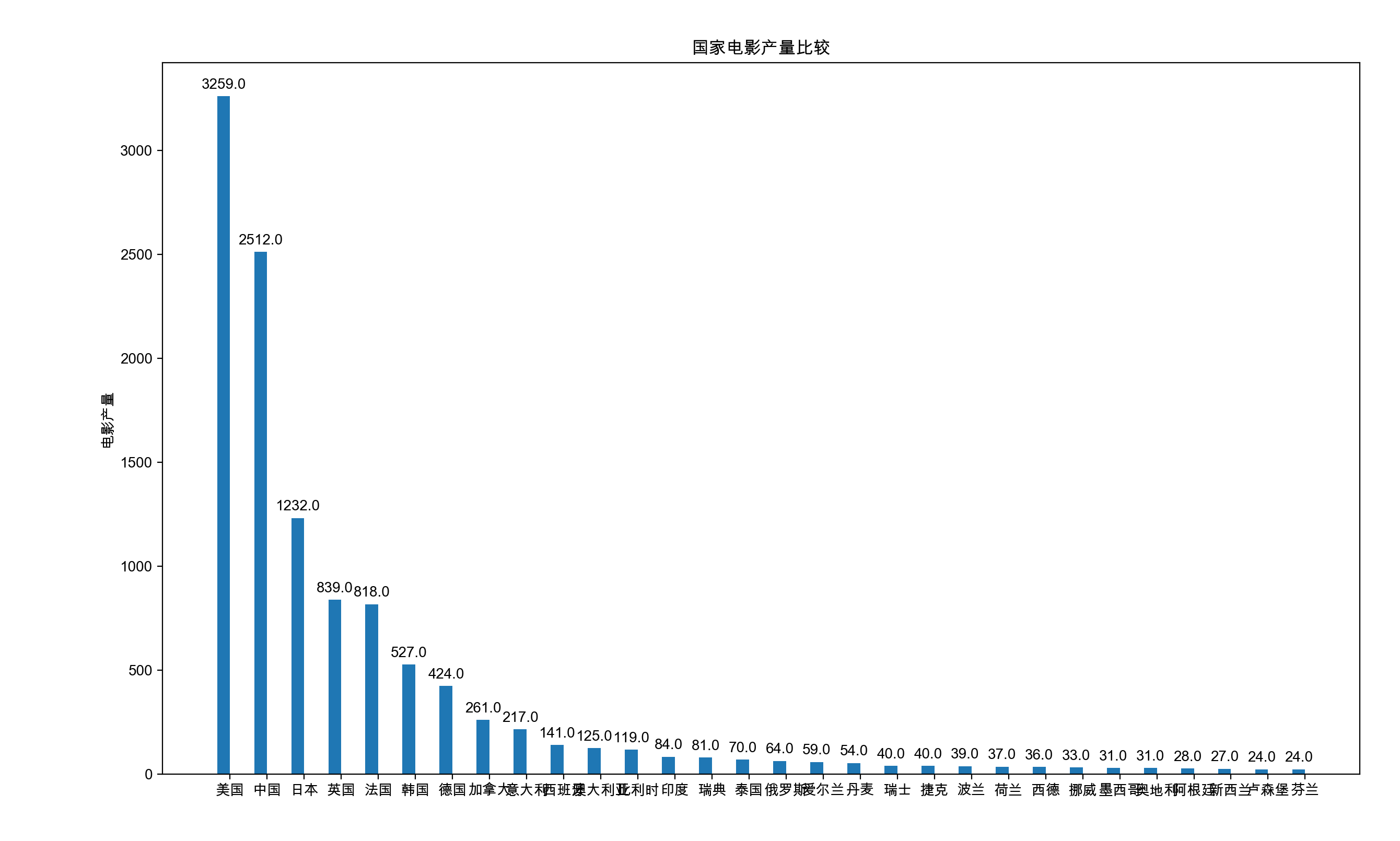
通过评分与观看人数，我们可以将电影分为高分热片（观看人数多，好评高，位于右上角区域）、冷门佳片（观看人数少，好评高，位于右下角区域）、烂片（观看人数少，差评多，位于左下角区域）这三类。

1. **评分与电影类型及数量的相关性**



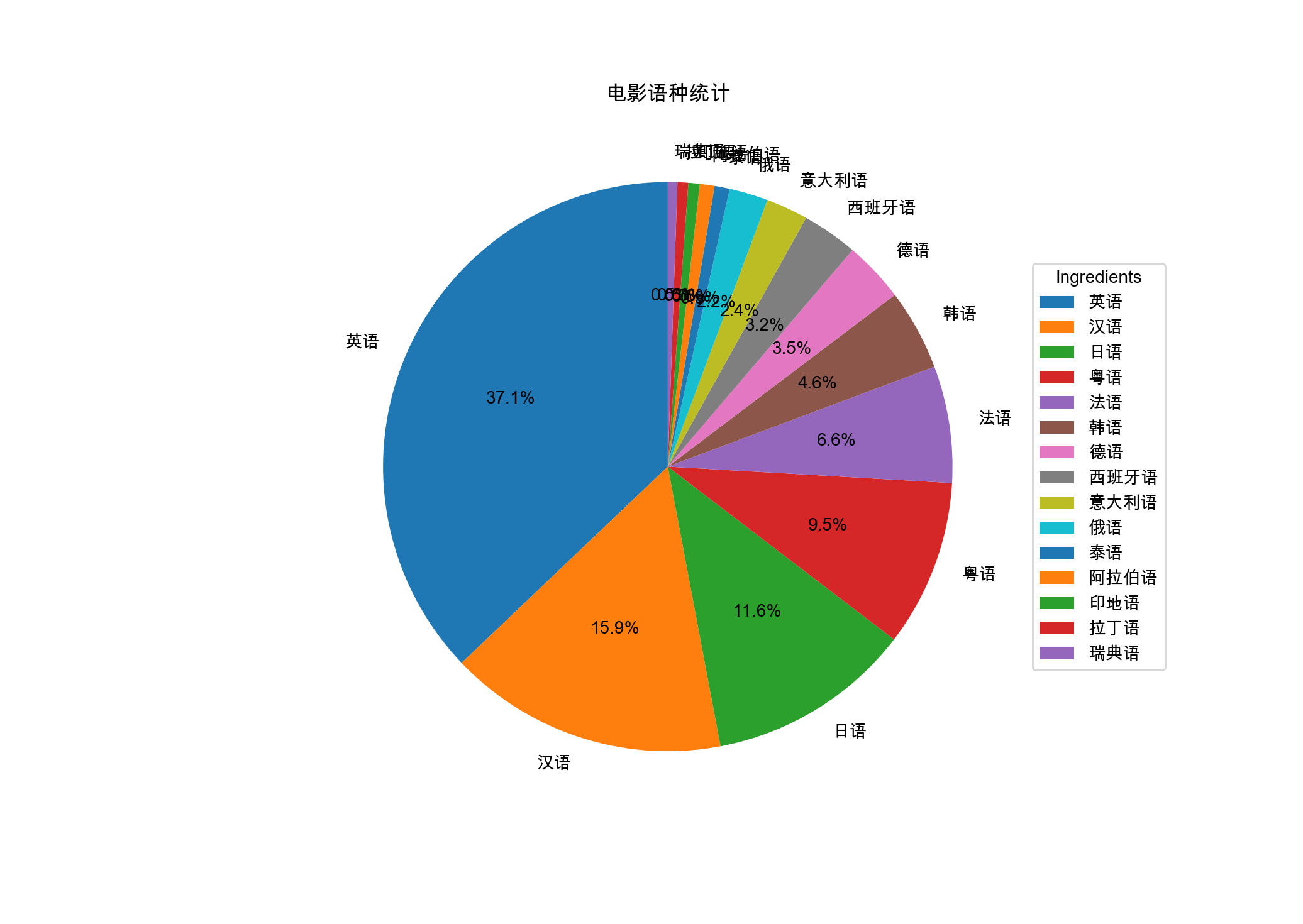
最受欢迎的电影类型为剧情（4661）、喜剧（2555）、动作（1962）、爱情（1884）、惊悚（1369）和犯罪（1229）这5大类，括号内为出现的频次。且主要的评价区域依然集中在中位偏上，也就是6至8分的位置。其中剧情类型的电影，好评占比最大。

1. **国家电影总数量分布**



美国电影数量最多，3259条，占总数据样本的36.7%，非常强势。其次为中日英法，头部效应非常明显。此外，电影数量少的原因也有很多，比如没有被我们关注到，导致没有被网站收录。也有可能电影所代表的的文娱产业，在经济高速发展的国家会远高于落后国家。具体原因需要再进行分析。

1. **电影语种统计**



排名前5的语种分别为英语、普通话、日语、粤语、法语。与各国电影总量成正比关系，基本在什么国家发行，用的就是什么语言。

1. **电影类型分布情况**



最受欢迎的电影类型为剧情（4661）、喜剧（2555）、动作（1962）、爱情（1884）、惊悚（1369）和犯罪（1229）。可进一步探索电影类型与评分的相关性。