**《Python语言程序设计》**

**综合项目报告**

**综合项目一**

**题目：IMDB影评数据分析**

**姓名：**

**班级：**

**学号：**

**2023-2024学年第二学期**

**一、实践目的**

1．使用numpy、matplotlib、pandas对数据进行分析；

2. 并能对数据分析结果进行图形化展示，为使用者提供分析结果。

**二、实践内容**

对所给的数据进行分析并进行可视化，可以使用折线图、柱状图、饼图、散点图等适合的图形化方式，要求分析内容包括：

1.统计IMDB评分各评分段的影片数。

2.各年票房变化的趋势。

3上榜IMDB的次数最多的导演的前5位、前10位。

4.影片上榜的次数和影片的评分的相关性。

5.票房和Mate评分的相关性。

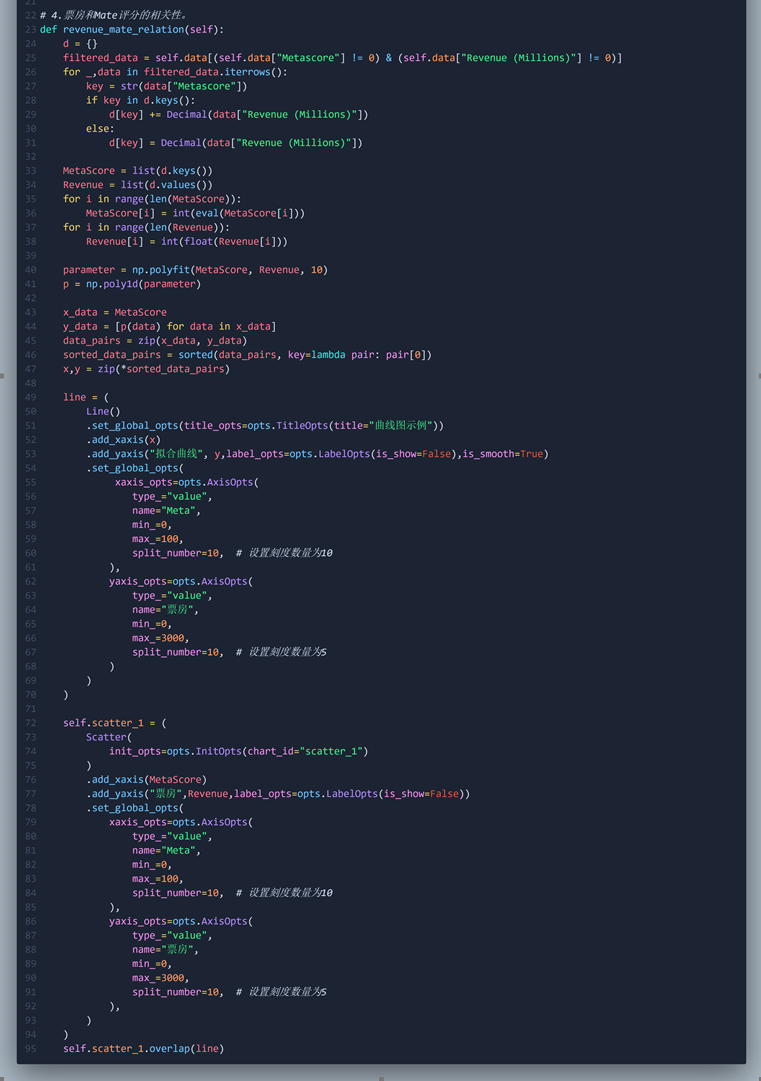
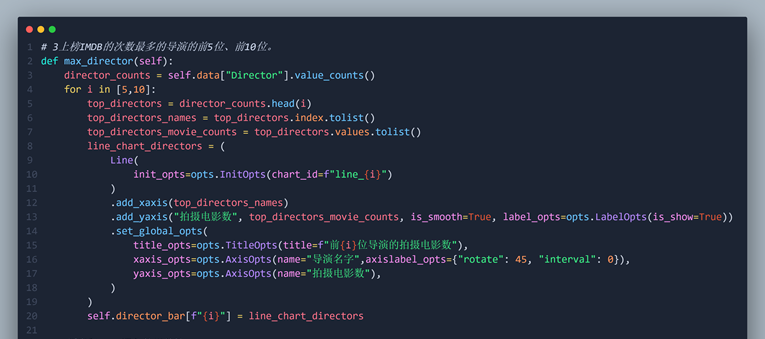
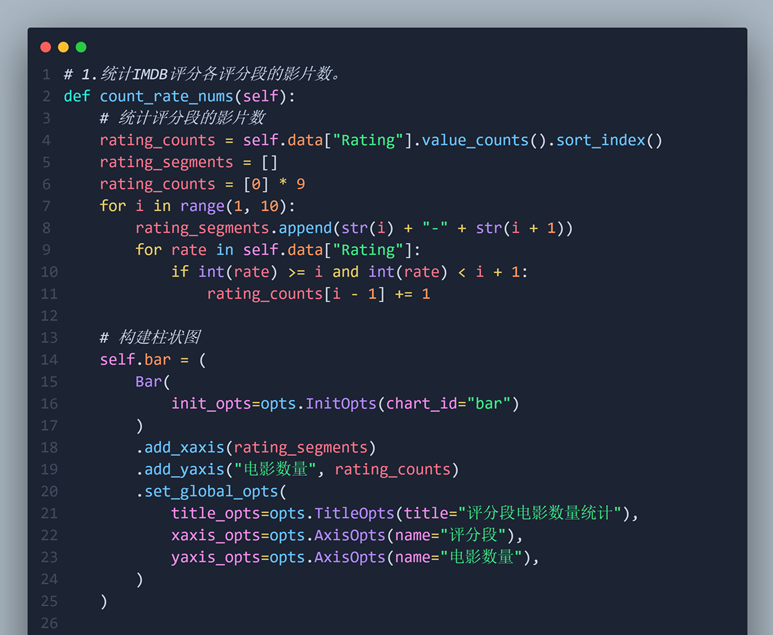
6.时长和Mate评分的相关性。

**三、实践步骤**

1. 算法描述

1. 首先，在Movie类的构造函数中，通过读取CSV文件，将电影数据加载到内存中，并对缺失值进行填充。
2. count\_rate\_nums()方法用于统计IMDB评分各评分段的影片数。它首先创建一个空的评分段列表和一个长度为9的评分计数列表。然后，通过遍历电影数据中的每个评分，将其分到相应的评分段中，并增加对应评分段的计数。最后，使用pyecharts库创建柱状图，横轴为评分段，纵轴为电影数量，展示评分段电影数量的统计结果。
3. revenue\_change()方法用于展示各年票房变化的趋势。它首先按年份对电影数据进行分组，并计算每年的总票房。然后，使用pyecharts库创建折线图，横轴为年份，纵轴为票房金额，展示年度票房统计结果。
4. max\_director()方法用于展示上榜IMDB次数最多的导演的前5位和前10位。它首先统计每个导演的电影数，并选择前5位和前10位导演。然后，使用pyecharts库创建折线图，横轴为导演名字，纵轴为拍摄电影数，展示前5位和前10位导演的拍摄电影数统计结果。
5. revenue\_mate\_relation()方法用于展示票房和Metascore评分的相关性。它首先筛选出票房和Metascore评分都不为0的电影数据，并计算它们的相关性。然后，使用pyecharts库创建散点图和拟合曲线，横轴为Metascore评分，纵轴为票房金额，展示票房和Metascore评分的相关性。
6. runtime\_mate\_relation()方法用于展示时长和Metascore评分的相关性。它首先筛选出时长和Metascore评分都不为0的电影数据，并计算它们的相关性。然后，使用pyecharts库创建散点图和拟合曲线，横轴为电影时长，纵轴为Metascore评分，展示时长和Metascore评分的相关性。
7. 最后，在run()方法中，依次调用以上各个方法，并使用pyecharts库创建一个可交互的页面，并将结果保存为HTML文件。

2. 代码描述



3. 调试过程

在使用pyecharts库进行渲染时，经常出现因为数据类型错误导致渲染失败，后来将数据都保留两位小数后成功渲染。

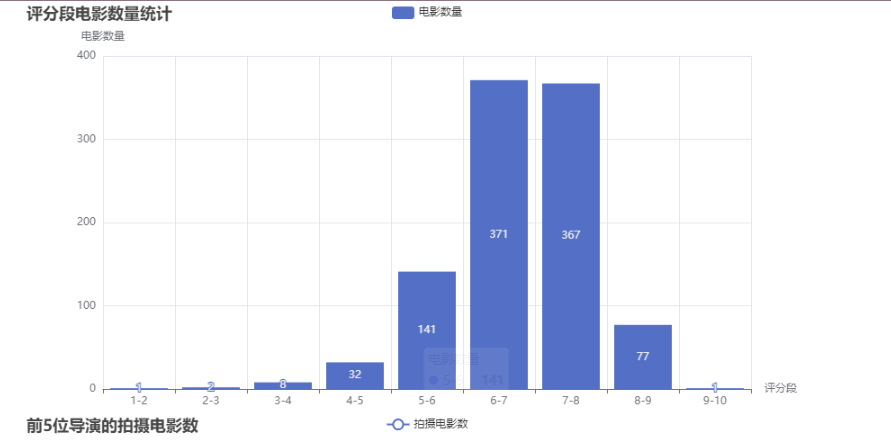
**四、结果及分析**

1. 测试数据

给定IDMB影评数据分析，IMDB-Movie-Data.csv文件

2. 实验结果

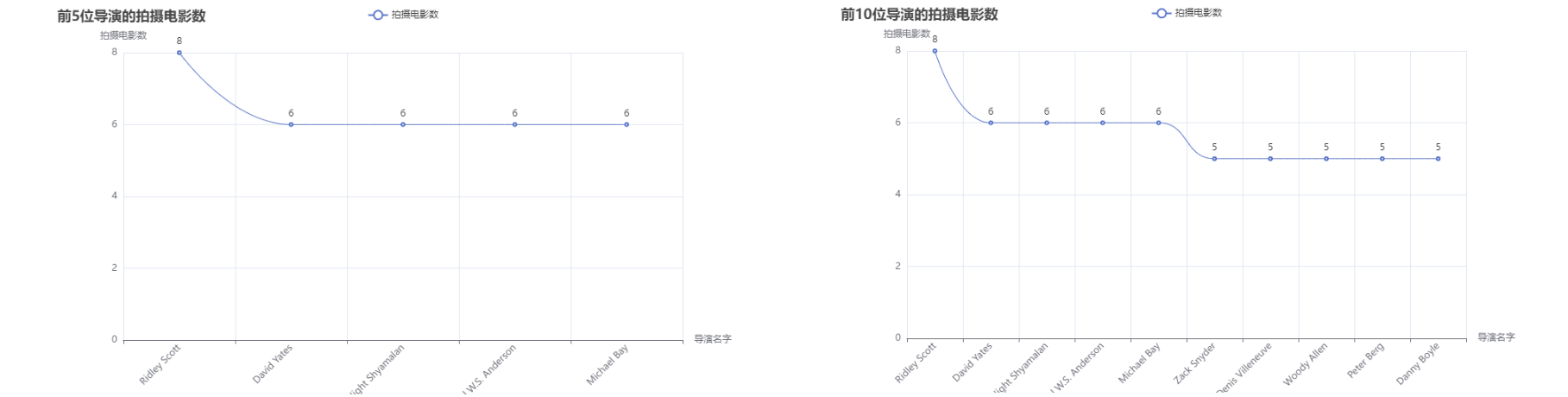
（1）统计IMDB评分各评分段的影片数。



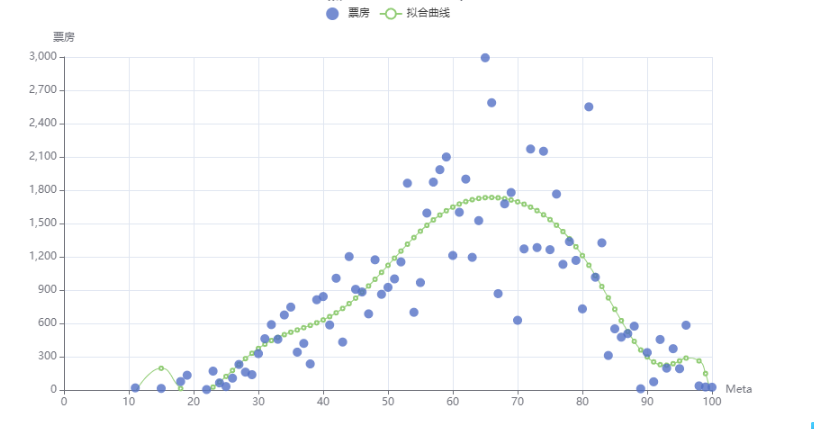
（2）各年票房变化的趋势。



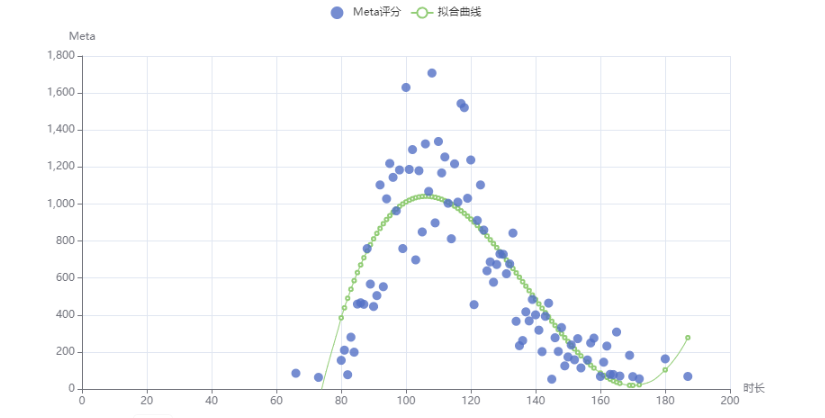
（3）上榜IMDB的次数最多的导演的前5位、前10位。



（4）票房和Mate评分的相关性。



(5) 时长和Mate评分的相关性。



3. 分析

（1） 由图可知，评分在6-8分之间的电影数量最多，分别6-7分371个，7-8分 367个，高评分和低评分数量都较少，整体曲线呈现一个抛物线。

（2） 由图可知2006-2016年票房的整体趋势是逐年上升的，票房数量稳步增长。

（3） 由图可知前五的导演为Ridley Scott上榜8次、David Yates上榜6次、M. Night Shyamalan上榜6次、Michael Bay上榜6次、Paul W.S. Anderson上榜6次

前十的导演为Ridley Scott上榜8次、David Yates上榜6次、M. Night Shyamalan上榜6次、Michael Bay上榜6次、Paul W.S. Anderson上榜6次、Antoine Fuqua上榜5次、Christopher Nolan上榜5次、Danny Boyle上榜5次、David Fincher上榜5次、Denis Villeneuve上榜5次。

（4）由图可知票房与Mate评分在以70分左右为界的低分段是正相关，而在高分段是负相关。

（5）由图可知时长与Mate评分在80分钟到110分钟左右是正相关，而在110分钟到170分钟是负相关。