

|  |
| --- |
| **计算机与信息工程学院** |

TMDB-5000-MovieAnalysis

课程名称： Python数据分析与应用

专 业：计算机科学与技术

学 号： 1710121265

姓 名： 吴清典

任课教师： 许涛

目录

[一.简介 3](#_Toc28367465)

[1.1数据集摘要 3](#_Toc28367466)

[1.2数据简介 4](#_Toc28367467)

[1.2.1 介绍数据集的概况 4](#_Toc28367468)

[1.2.2 描述分析目标 4](#_Toc28367469)

[1.2.3分析手段和方法 5](#_Toc28367470)

[1.2.4阐述结论 5](#_Toc28367471)

[二 数据处理 5](#_Toc28367472)

[2.1数据处理环境 5](#_Toc28367473)

[2.2 数据读取 6](#_Toc28367474)

[2.3 数据清洗 9](#_Toc28367475)

[三 数据可视化 11](#_Toc28367476)

[3.1问题一 11](#_Toc28367477)

[3.2问题二 14](#_Toc28367478)

[3.3问题三 17](#_Toc28367479)

[3.4问题四 20](#_Toc28367480)

[3.5问题五 22](#_Toc28367481)

[3.6问题六 23](#_Toc28367482)

[3.7问题七 24](#_Toc28367483)

[四结论和展望 26](#_Toc28367484)

[4.1总结 26](#_Toc28367485)

[4.2展望 27](#_Toc28367486)

# 一.简介

## 1.1数据集摘要

|  |  |
| --- | --- |
| 名称 | TMDB 5000 Movie Dataset（TMDB 5000电影数据集） |
| 特征简介 | 'budget'， 'genres'，'homepage'，'id'  'keywords'，'original\_language'  'original\_title'，'overview'  'popularity'，'production\_companies'  'production\_countries'，'release\_date'  'revenue'，'runtime'，'spoken\_languages'  'status'，'tagline'，'title'  'vote\_average'，'vote\_count'  'movie\_id'，'title'，'cast'，'crew' |
| 记录数 | 4813 |
| 分析目标 | 问题1：电影风格随时间的变化趋势  问题2：不同风格电影的收益能力和年份与什么类型的电影风格最相关  问题3：比较行业内Universal Pictures与Paramount Pictures两家巨头公司的业绩和各自的电影风格  问题4：票房收入与哪些因素最相关以及电影的关键字  问题5：各种电影类型所占的比例各是多少，其中占比最大的是什么类型  问题6：如何实现电影评分预测  问题7：如何使用Kmeans聚类对数据进行数据分析 |
| 分析思路及方法 | 1. 提出要分析的问题 2. 浏览数据，理解数据 3. 数据清洗 4. 建立模型 5. 数据可视化 6. 对建模可视化后的数据形成数据报告分析 |

## 1.2数据简介

### 1.2.1 介绍数据集的概况

（1）数据来源：<https://www.kaggle.com/tmdb/tmdb-movie-metadata>

（2）数据属性：（如下配上一张属性的表）

**表1.2.1 数据列名**



（3）数据量：一共两个表格，每个表格中各有4813条数据

（4）基本统计特征：tmdb\_5000\_movies.csv表中共有20个变量，tmdb\_5000\_credits表中共有4个变量名。

### 1.2.2 描述分析目标

对于数据提出了四个问题（如下）：

问题1：电影风格随时间的变化趋势

问题2：不同风格电影的收益能力和年份与什么类型的电影风格最相关

问题3：比较行业内Universal Pictures与Paramount Pictures两家巨头公司的业绩

问题4：票房收入与哪些因素最相关以及电影的关键字

问题5：各种电影类型所占的比例各是多少，其中占比最大的是什么类型。

问题6：如何实现电影评分预测

### 1.2.3分析手段和方法

1.将下载后的数据读取出来后，理解数据后对数据进行处理。

2.数据清洗（根据问题对数据进行一定程度上的修改）

3.建立模型

4.数据可视化

5.形成数据分析报告

### 1.2.4阐述结论

结论1：从上世纪90年代开始，整个电影市场呈现爆发式增长。其中，排名前五的戏剧类（Drama）、喜剧类（Comedy）、惊悚类（Thriller）、浪漫类（Romance）、动作类（Action）电影数量增长显著，排名前六至前九的类型增长相对比较缓慢。

结论2：从图中可以看出不同类型的电影收益状况不同，其中又以动画类（Animation）和历史类（History）收益最高。年份与（Western）类型相关性最高，比其他类型的高出不少。

结论3：Universal Pictrues总票房收入高于Paramount Pictures，两公司排名前九的电影风格通过饼图展示出来，两公司的电影份风格大致相同相差不多。

结论4：电影评次与票房收入最相关，电影预算与票房收入高度相关，受欢迎度与评次高度相关，其他的因素与票房收入相关性相对较低。

结论5：各种电影类型的占比比例如图所示，其中占比类型最大的是Drama。

结论6：根据影片类型、导演和主演，对于待预测的影片，筛选出这3个因素与之相似程度最高的5部影片，计算它们的平均评分，作为待预测影片的评分。

结论7：从数据集中挑选出‘popularity’和‘vote\_count’数据使用Kmeans聚类来进行分析，分析结果可以根据图来显示出来。

# 二 数据处理

## 2.1数据处理环境

环境：win10+Anaconda3+Spyder+python3.6

## 2.2 数据读取

1.先导入需要的包

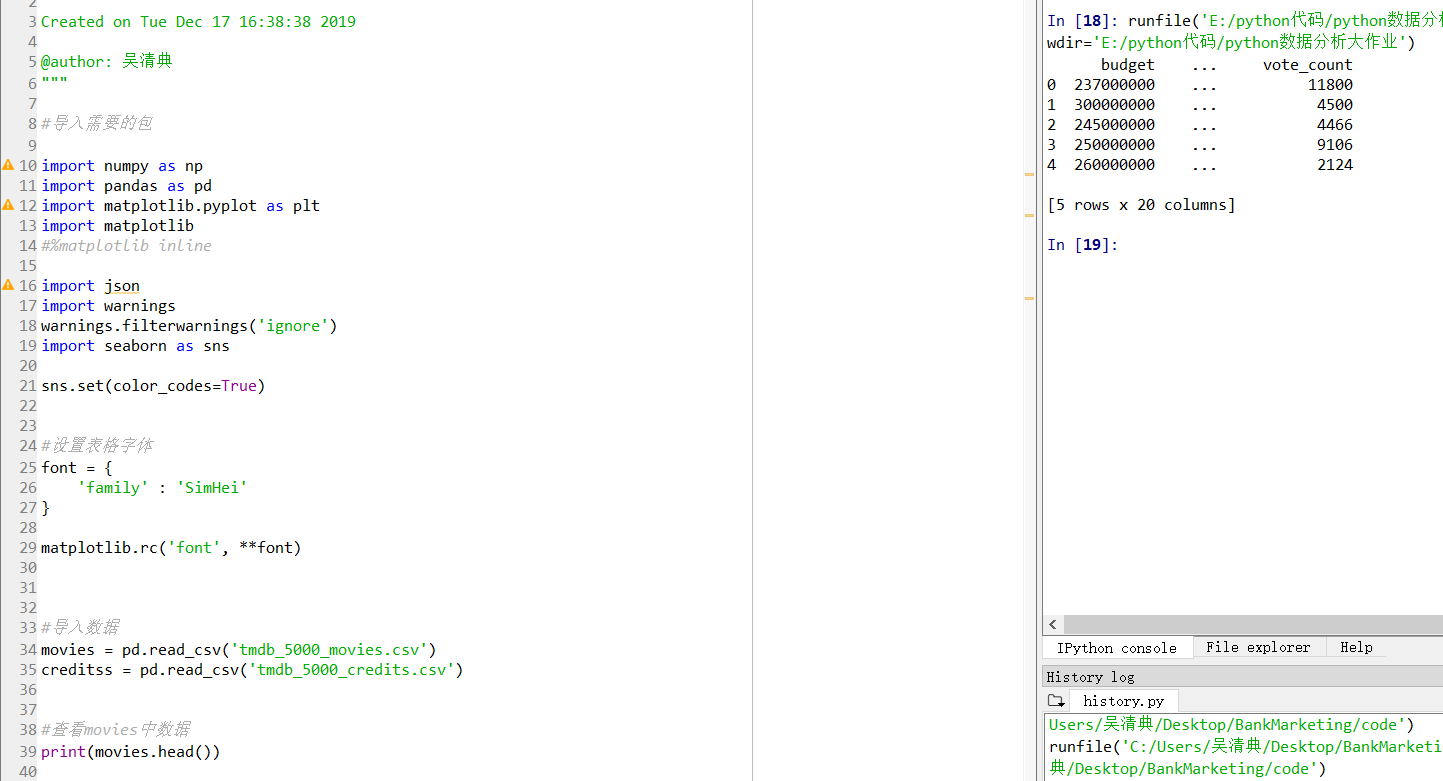


图2.2.1 导入需要的包图

2.读取，查看数据

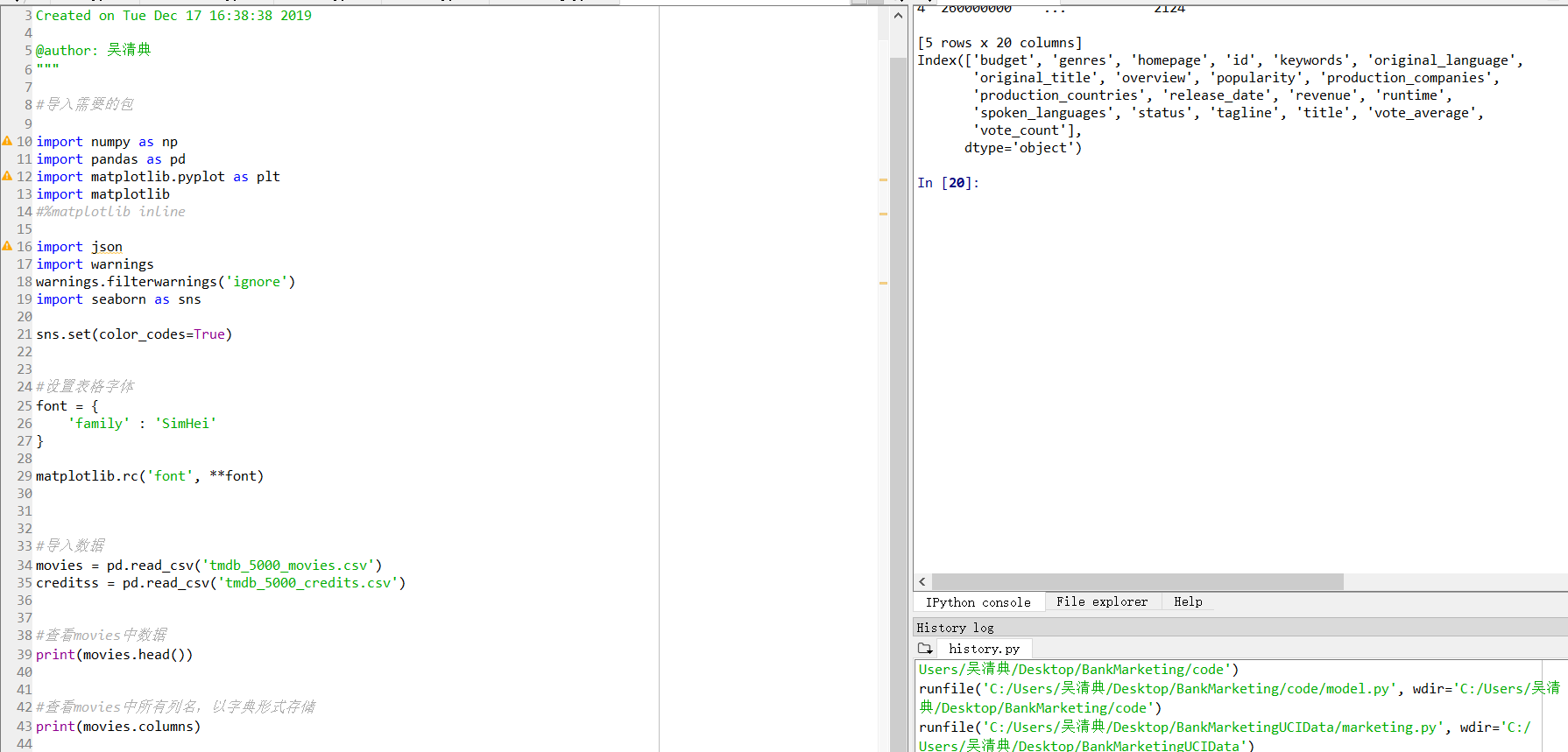
（1）使用pd.read\_csv(‘data’)来读取数据

（2）打印print（movies.head（））大致查看一下数据 print(movies.head())



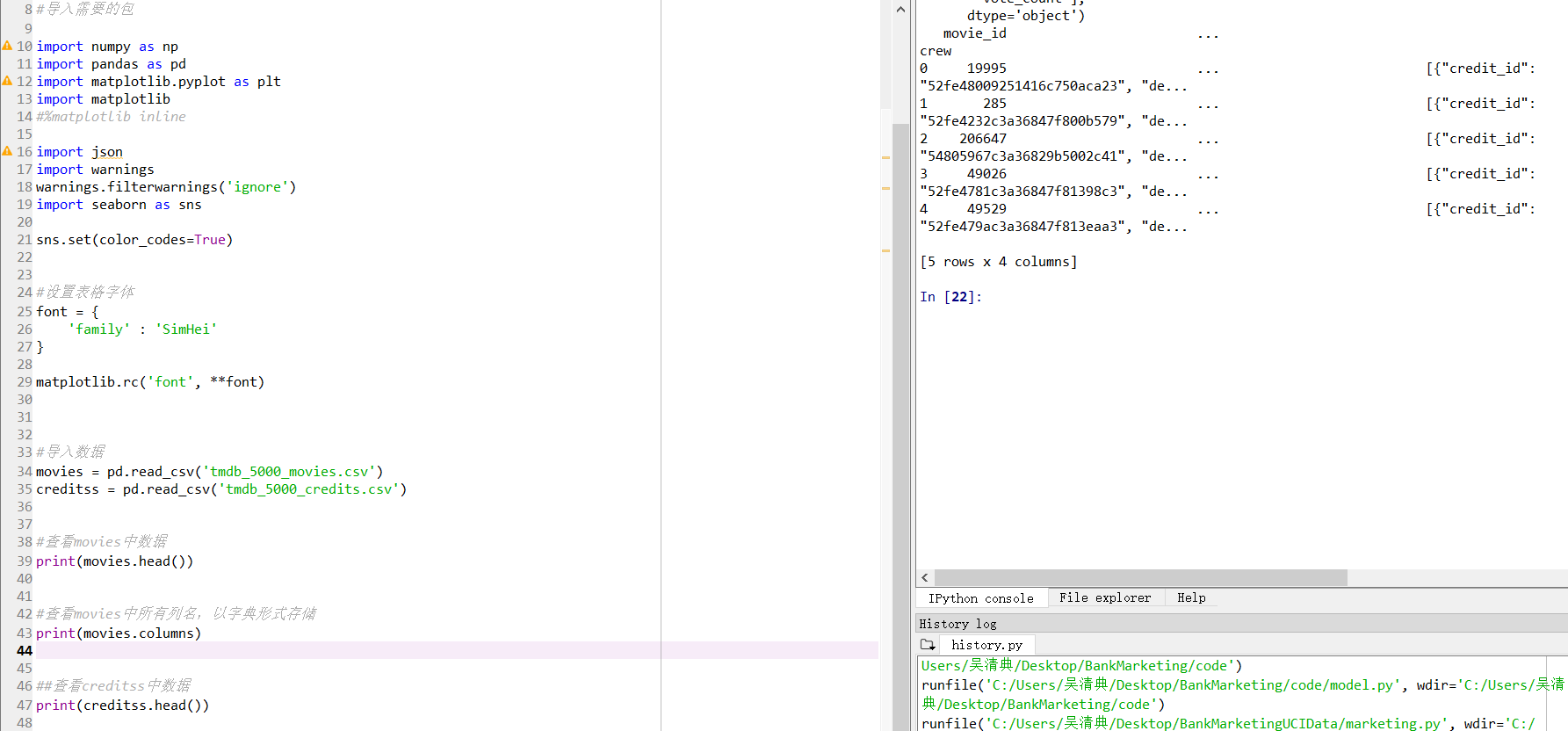
**图2.2.2 查看movies的数据图**

（3）查看一下movies数据的列名 print(movies.columns)



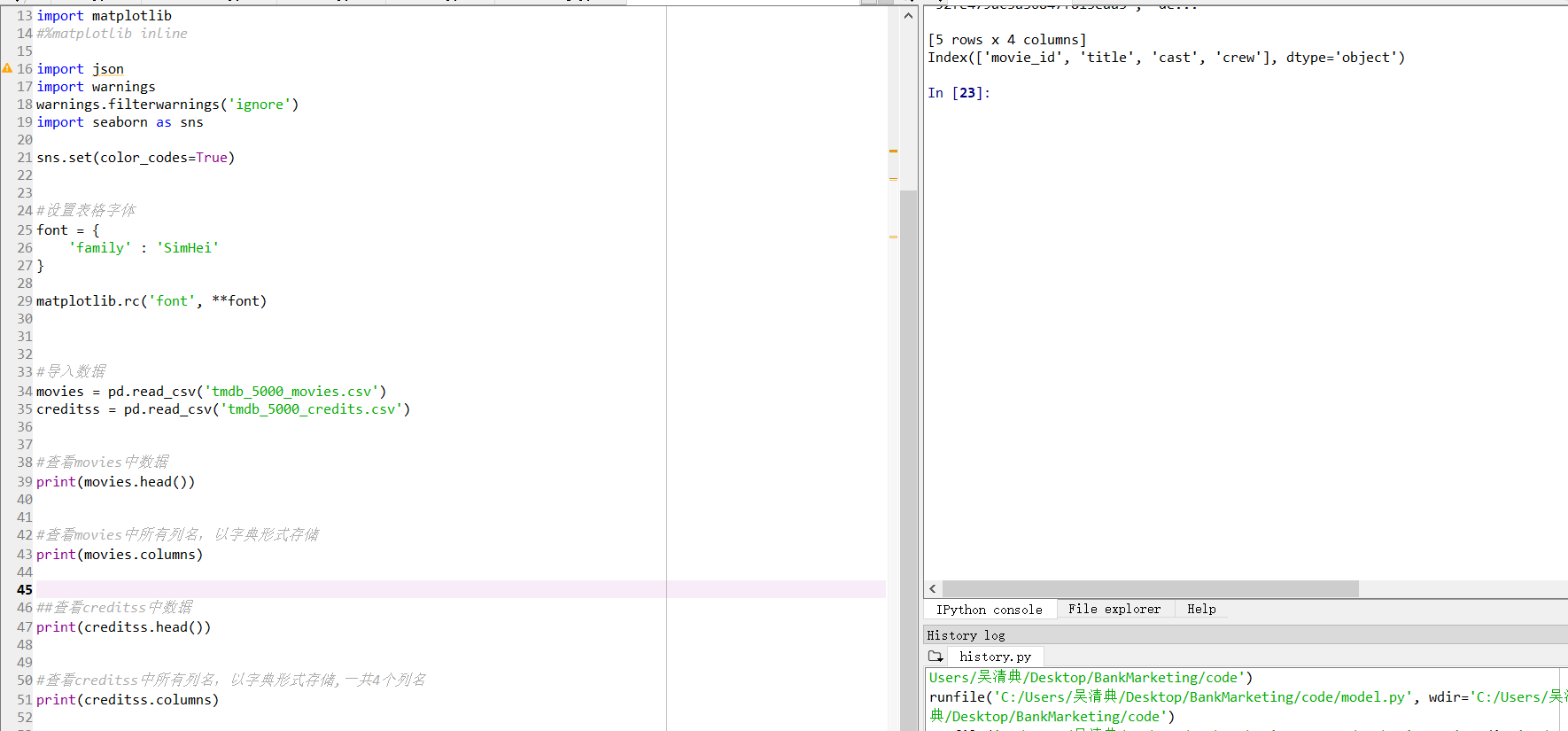
**图2.2.3查看movies数据的列名图**

（4）查看一下creditss数据的前五条数据 print(credits.head())

****

**图2.2.4 查看creditss的数据图**

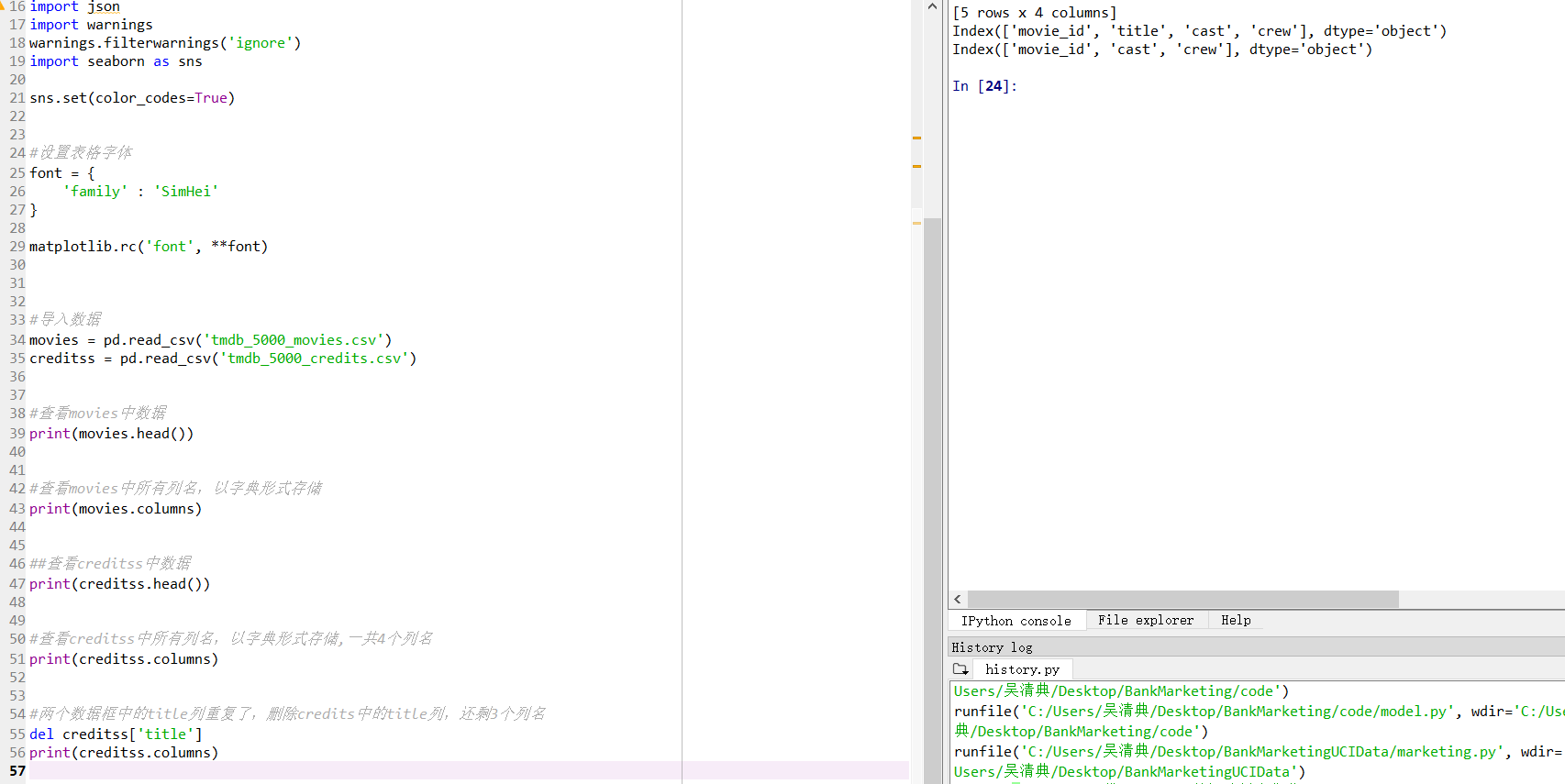
（5）查看一下creditss数据的列名 Print(Credits.columns)

****

**图2.2.5查看creditss数据的列名图**

（6）因为两个数据框中的title列重复，删除了credits中的title列

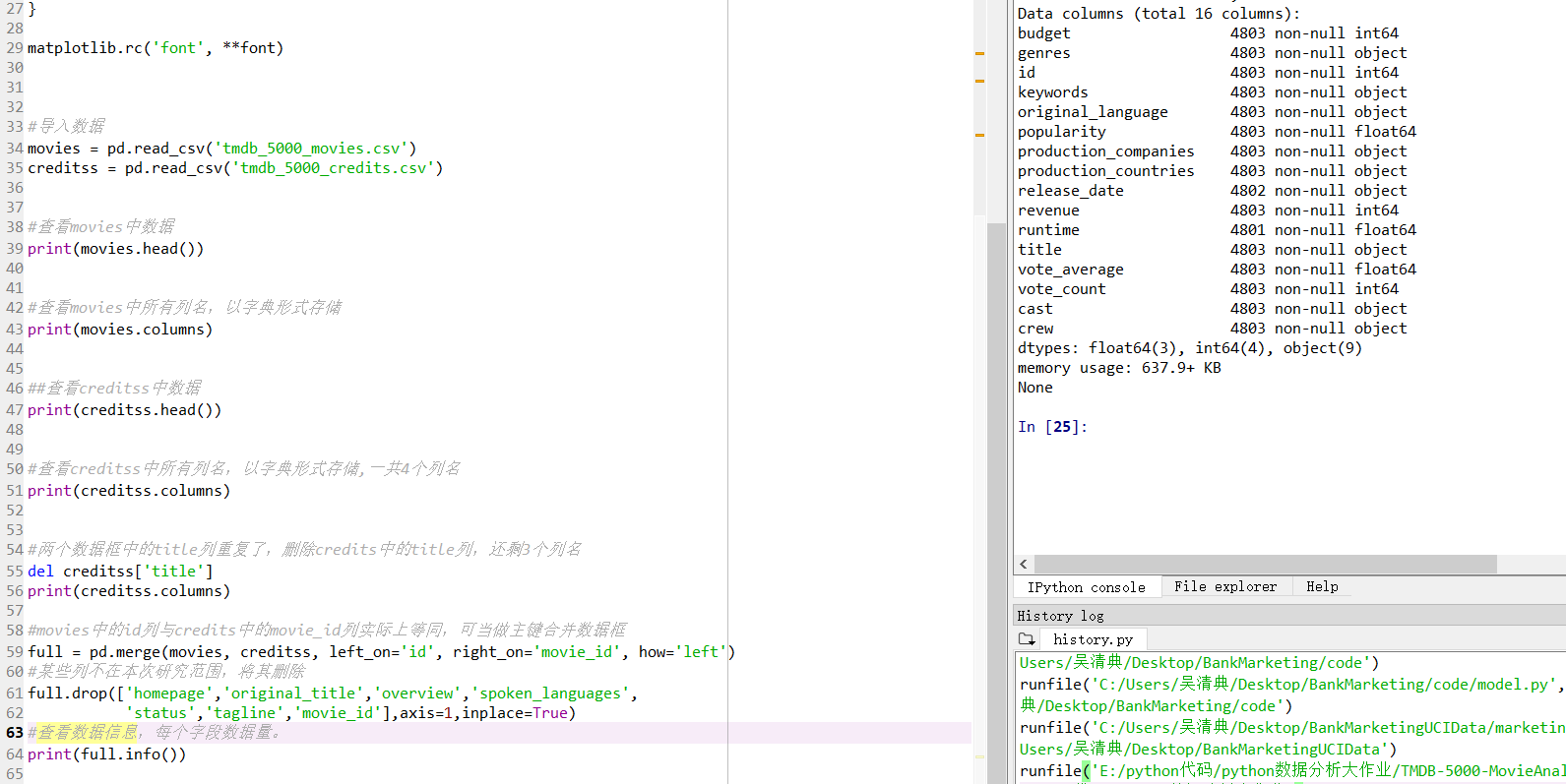
del creditss['title']



**图2.2.6删除重复title列图**

（7）因为movies中的id和credits中的movie\_id列等同，所以合并数据，合并后根据需要提出的要求选择需要的数据并删掉一些不需要的数据，最后查看数据信息，字段数据量。

删除：Del credits[‘title’]，合并：pd.merge(),删除不需要的数据：full.drop（）



**图2.2.7查看full数据信息图**

## 2.3 数据清洗

（1）查看合并后的数据哪有缺失值，找出缺失处并补充缺失值

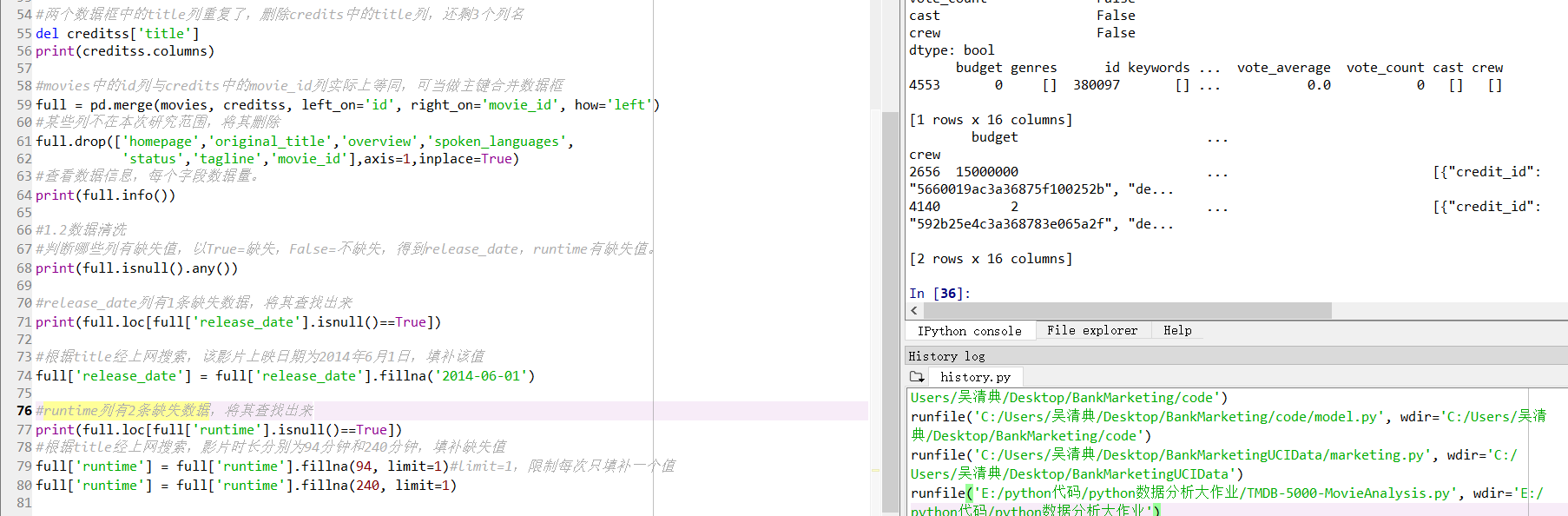
查找缺失：full.isnull().any()

填充数据：

full['release\_date'] = full['release\_date'].fillna('2014-06-01')

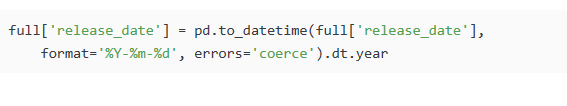
full['runtime'] = full['runtime'].fillna(94, limit=1)#limit=1，限制每次只填补一个值

full['runtime'] = full['runtime'].fillna(240, limit=1)



**图2.3.1查找缺失，填充数据图**

（2）将release\_date列转换为日期类型pd.to\_datetime()



**图2.3.2转化为时间类型图**

（3）因为数据集中有json数据，所以需要解析json数据。



**图2.3.3解析json数据图**

（4）根据所要分析的问题取出一些主演和导演的数据。



**图2.3.4取出导演，主演数据图**

# 三 数据可视化

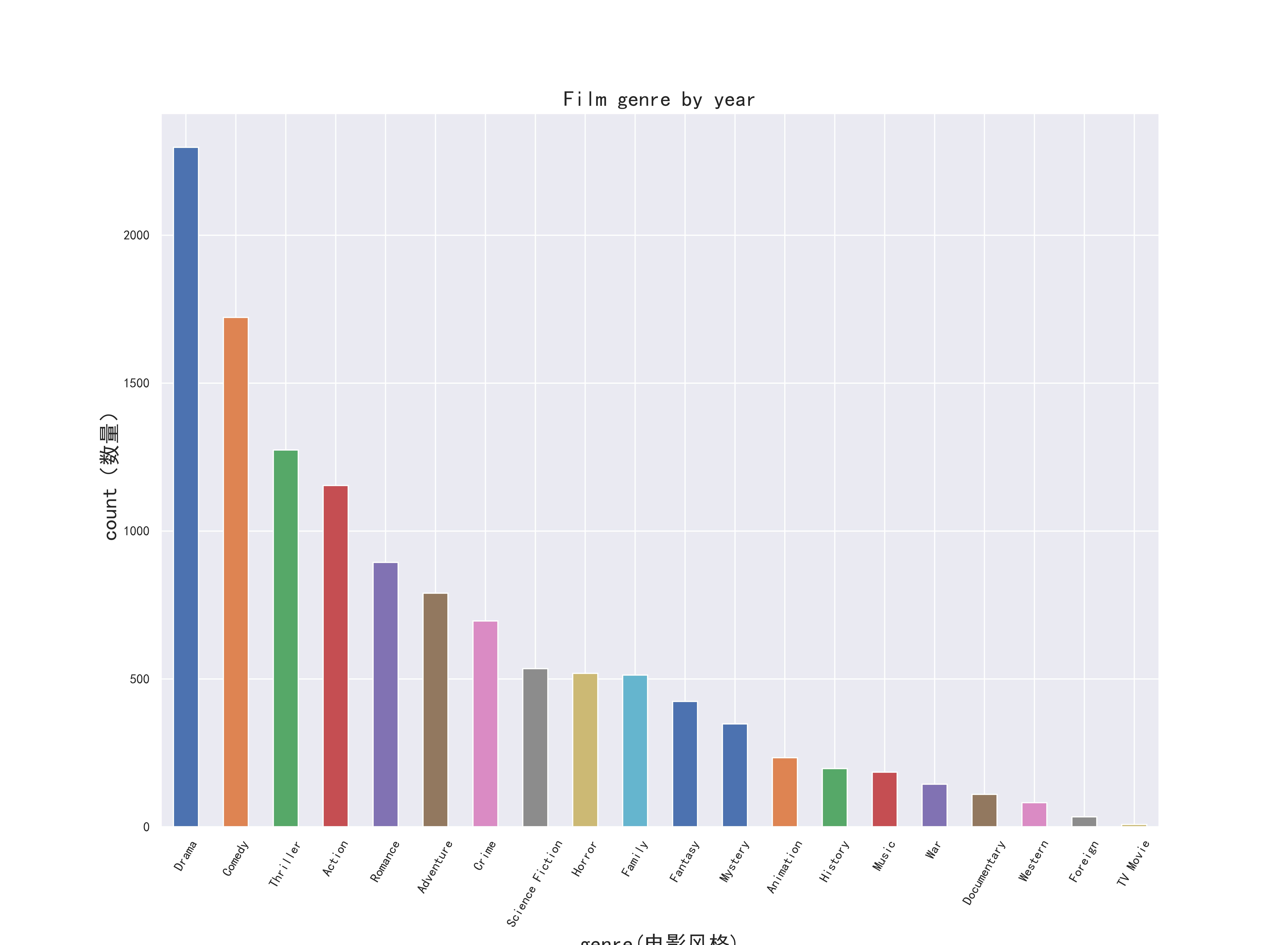
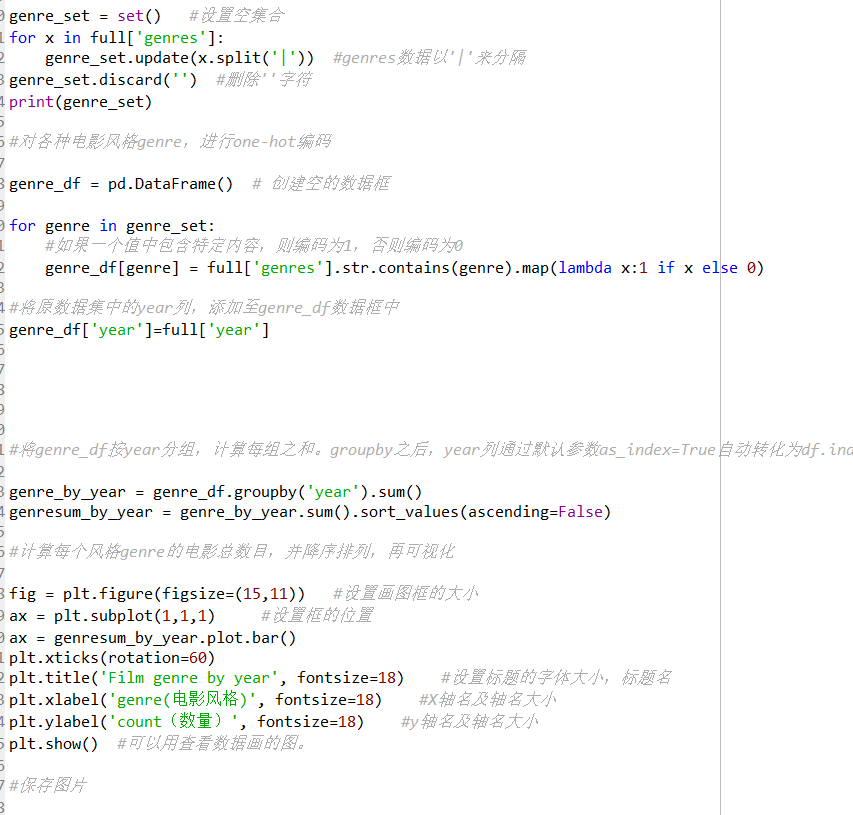
### 3.1问题一

问题1：电影风格随时间的变化趋势。

解决方法：

（1）先将所有的电影风格取出并去重，将年份的列取出，将风格属性按年份分组计算每组的和并降序排序，可视化出电影风格的数量。

**图3.1.1电影风格数量代码图**

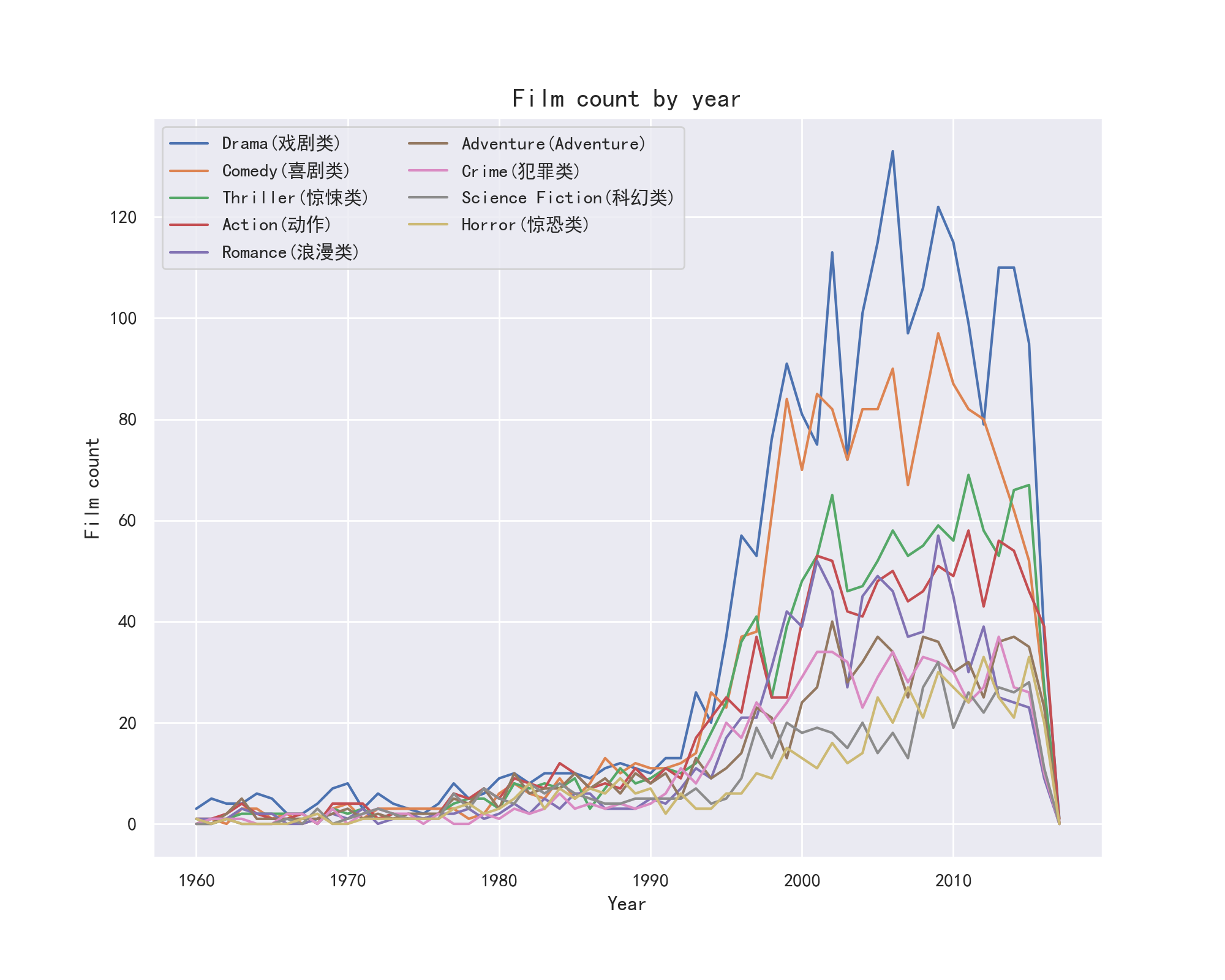


**图3.1.2电影风格数量图**

（2）先将所有的电影风格TOP9筛选出来，然后可视化电影风格随时间变化的趋势（1960-2010年）



**图3.1.3TOP9File by year代码图**



**图3.1.4TOP9File by year图**

问题一结论：从上世纪90年代开始，整个电影市场呈现爆发式增长。其中，排名前五的戏剧类（Drama）、喜剧类（Comedy）、惊悚类（Thriller）、浪漫类（Romance）、动作类（Action）电影数量增长显著，排名前六至前九的类型增长相对比较缓慢。

### 3.2问题二

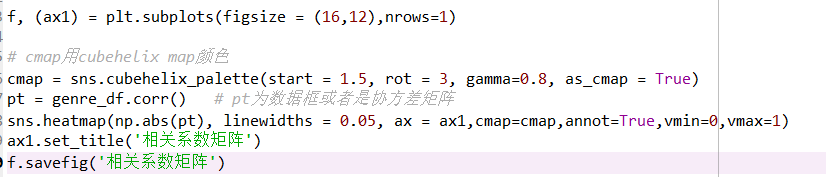
问题2：（1）不同风格电影的收益能力。

（2）年份与什么类型的电影风格最相关。

解决方法：先增加收益数据列和收益率列，根据画出对应数据的相关系数矩阵热力图、直方图和折线图来进行分析。



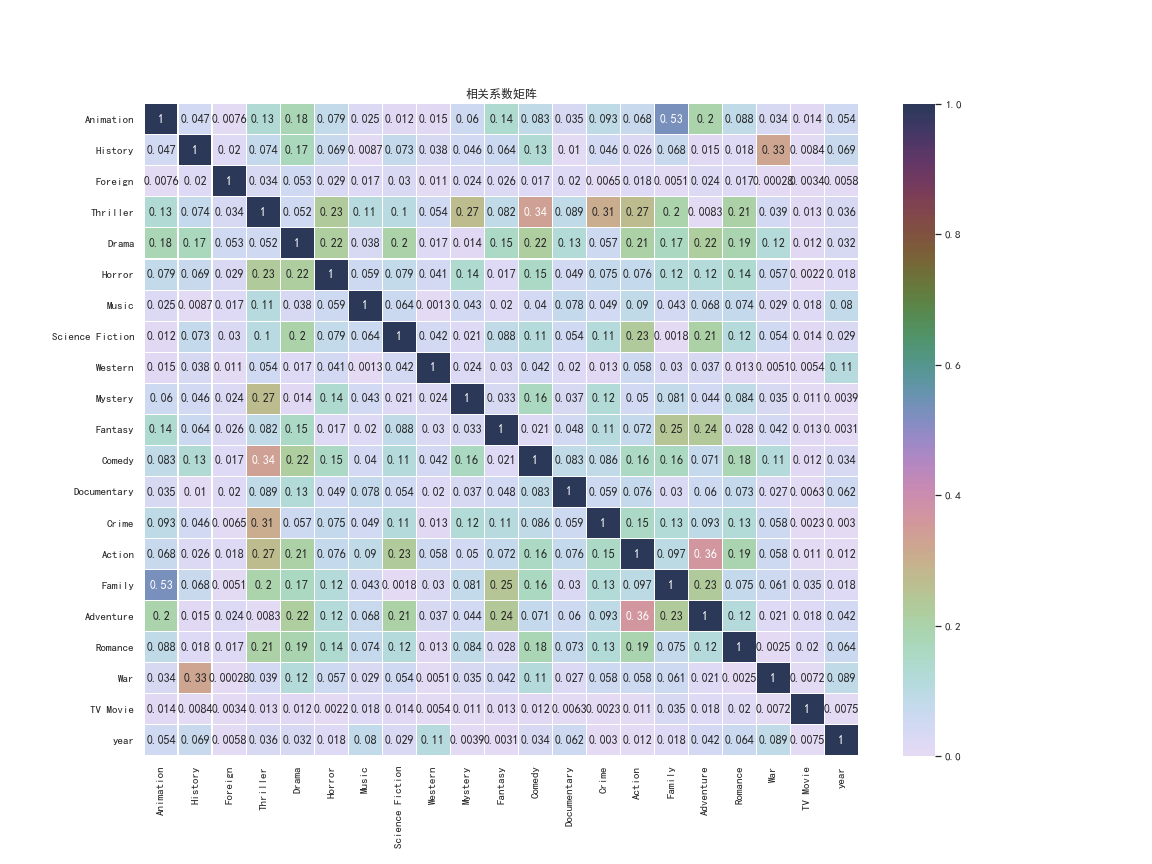
**图3.2.1增加收益列和收益率列代码图**



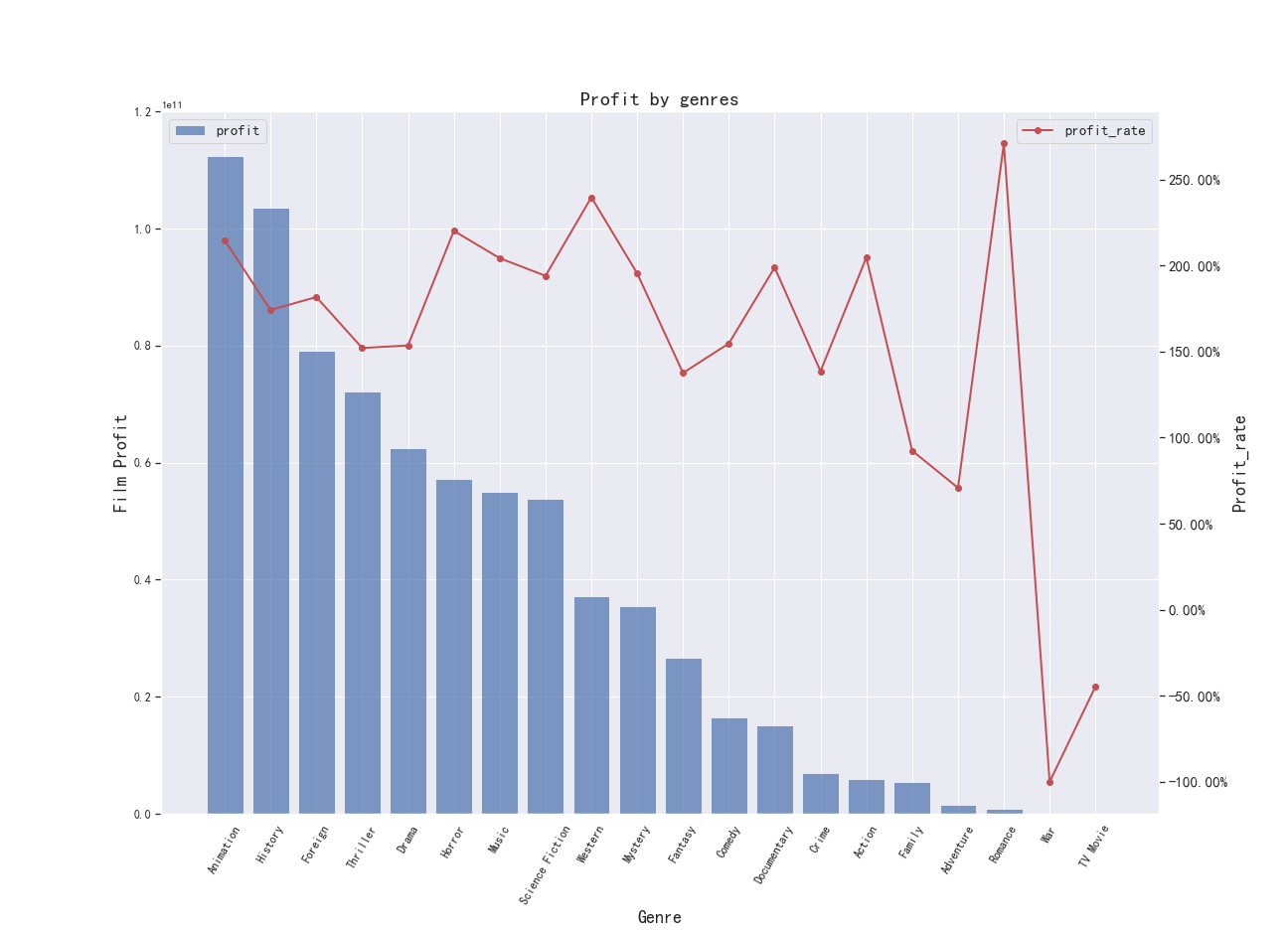
**图3.2.2相关系数矩阵热力代码图**



**图3.2.3可视化收益列和收益率列代码图**



**图3.2.4相关系数矩阵热力图**



**图3.2.5收益列和收益率图**

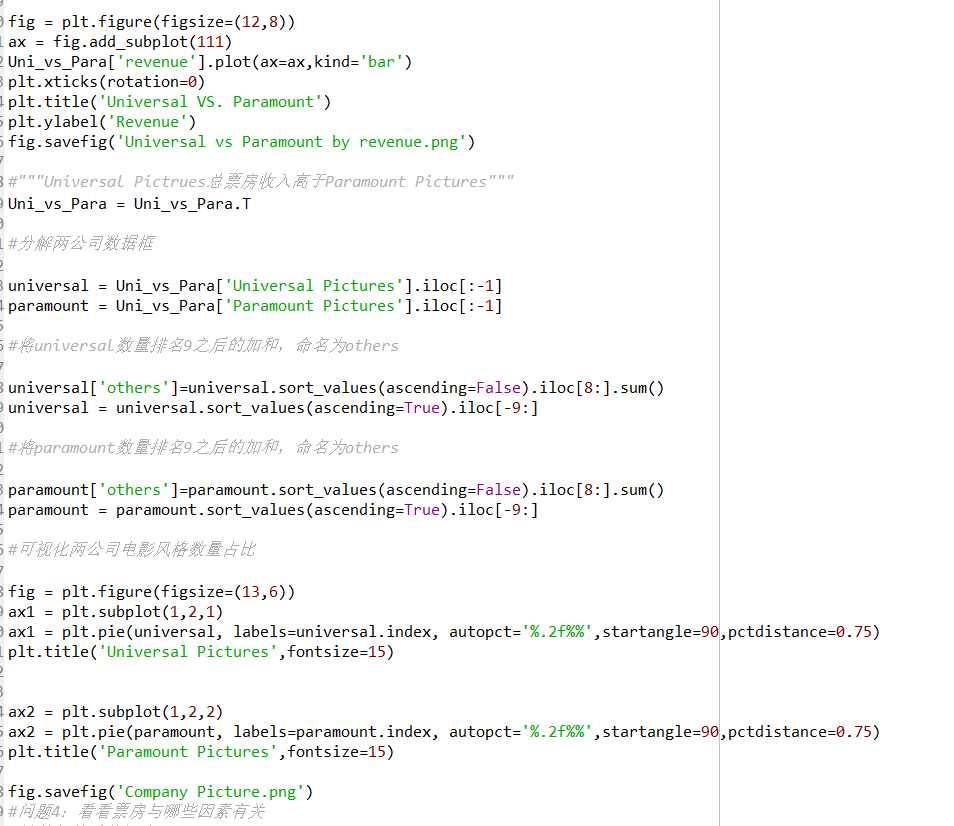
问题二结论：（1）从图中可以看出不同类型的电影收益状况不同，其中又以动画类（Animation）和历史类（History）收益最高。

（2）年份与（Western）类型相关性最高，比其他类型的高出不少。

### 3.3问题三

问题3：比较行业内Universal Pictures与Paramount Pictures两家巨头公司的业绩

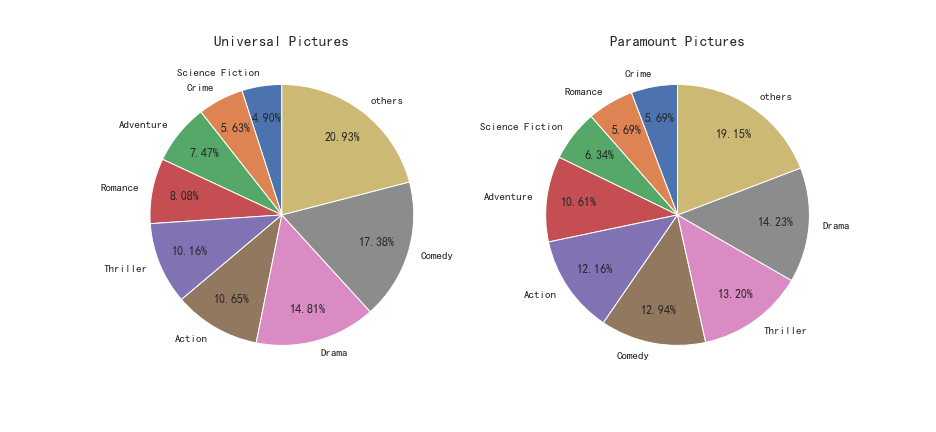
解决方法： 可视化两公司电影风格数量占比，创建两家公司的数据框，计算两家公司各自收益总额，可视化两家公司票房的收入对比。



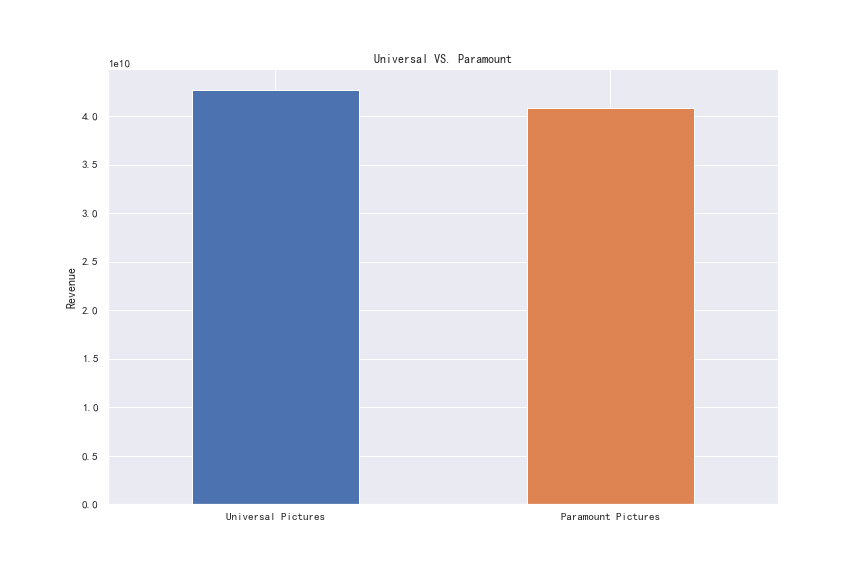
**图3.3.1两公司电影风格比例代码图**



**图3.3.2两公司票房收入对比代码图**



**图3.3.3两公司电影风格比例图**



**图3.3.4两公司票房收入对比图**

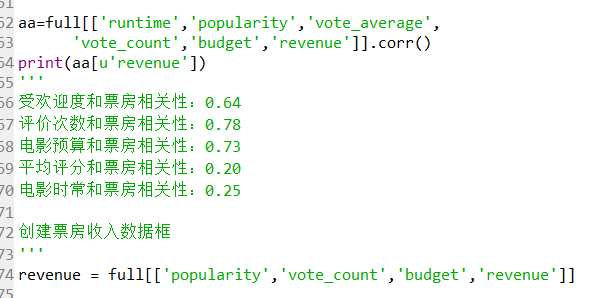
问题三结论：从图中可以看出Universal Pictrues总票房收入高于Paramount Pictures，两公司排名前九的电影风格通过饼图展示出来，两公司的电影份风格大致相同相差不多。

### 3.4问题四

问题4：票房收入与哪些因素最相关。

解决方法：

通常来说影响票房的收入（Revenue）因素分别有电影时长（Runtime）、欢迎程度（Popularity）、平均评分（Vote\_average）、评价次数（Vote\_count）和电影预算（Budget），所以通过计算相关系数矩阵来算出各种因素和票房的相关性，然后可视化相关性较高的散点图并配上线性回归线以便更好的观察。



**图3.4.1票房与其他属性的相关性图**

通过计算可以分别算出：

受欢迎度和票房相关性：0.64

评价次数和票房相关性：0.78

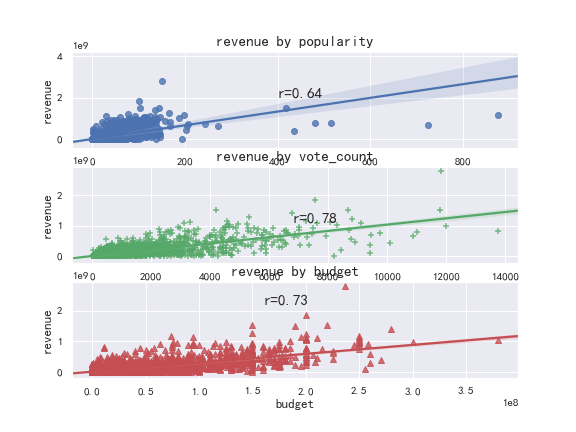
电影预算和票房相关性：0.73

平均评分和票房相关性：0.20

电影时常和票房相关性：0.25

因为平均评分和电影时常相关性较低，在下图的可视化中就暂不可视化这两个属性。

**图3.4.2票房收入相关性散点代码图**



**图3.4.3票房收入相关性散点图**



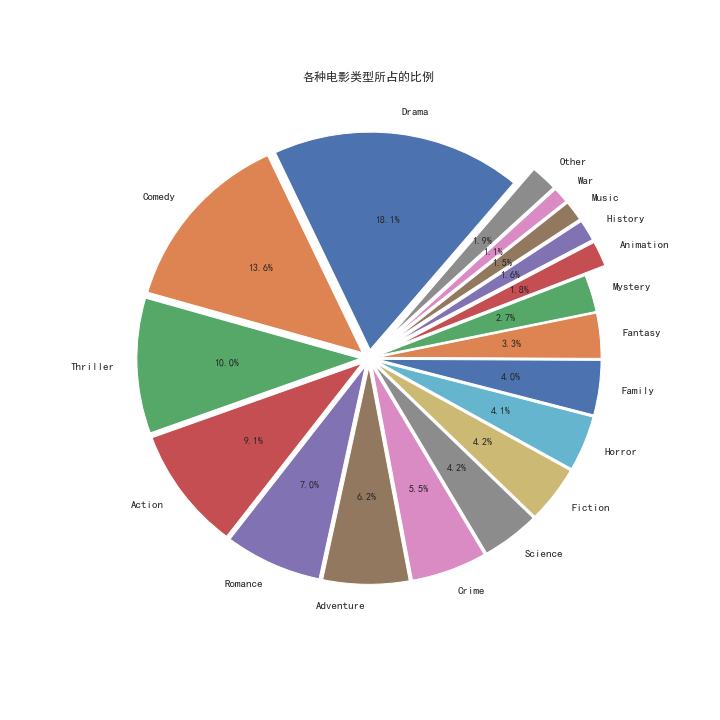
**图3.4.4关键字图**

问题四结论：经过计算相关系数矩阵和观察相关性散点图可以得出结论，在五大属性中：电影评次与票房收入最相关，电影预算与票房收入高度相关，受欢迎度与评次高度相关，其他的因素与票房收入相关性相对较低。

### 3.5问题五

问题5：各种电影类型所占的比例各是多少，其中占比最大的是什么类型。

解决方法：通过计算各种电影类型数量与总电影类型数量之比，通过饼图来显示出来各种类型的占比比例。



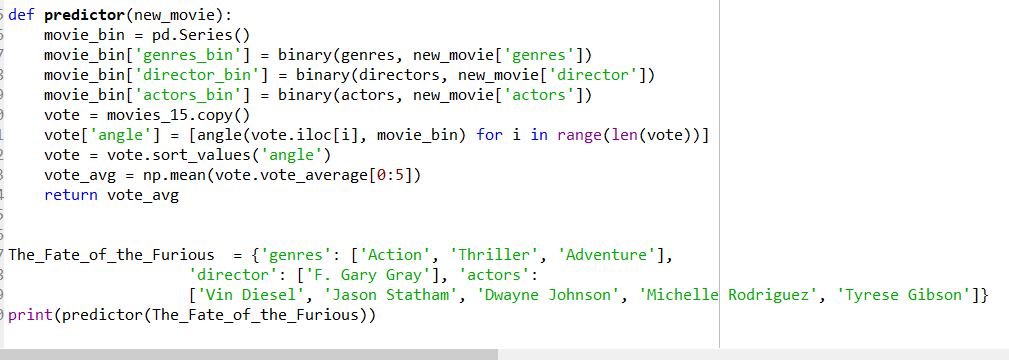
**图3.5.1各种电影类型占比比例图**

问题五结论：各种电影类型的占比比例如图所示，其中占比类型最大的是Drama，占到了百分之18.1的比例，与第二名（comedy类型）拉开了百分之4.5的差距，可见电影类型中数量最多的就是Drama类型。

### 3.6问题六

问题6：如何实现电影评分预测

解决方法：建议一个简单的预测模型，根据影片类型、导演和主演，对于待预测的影片，筛选出这3个因素与之相似程度最高的5部影片，计算它们的平均评分，作为待预测影片的评分。

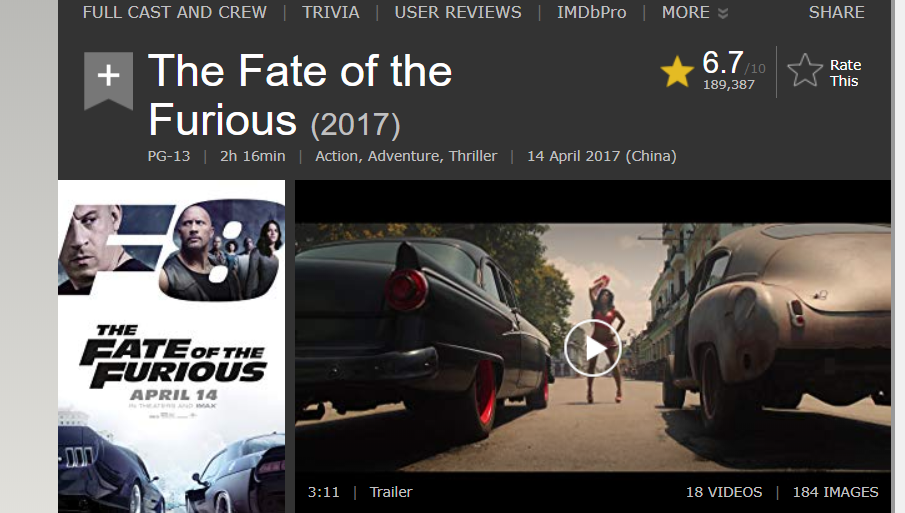


**图3.6.1电影评分预测代码图**

根据评分预测模型，在这里预测了近年上映的电影：速度与激情8（The Fate of the Furious）,我们预测的分数是6.6



**图3.6.2电影评分预测图**



**图3.6.3电影评分图**

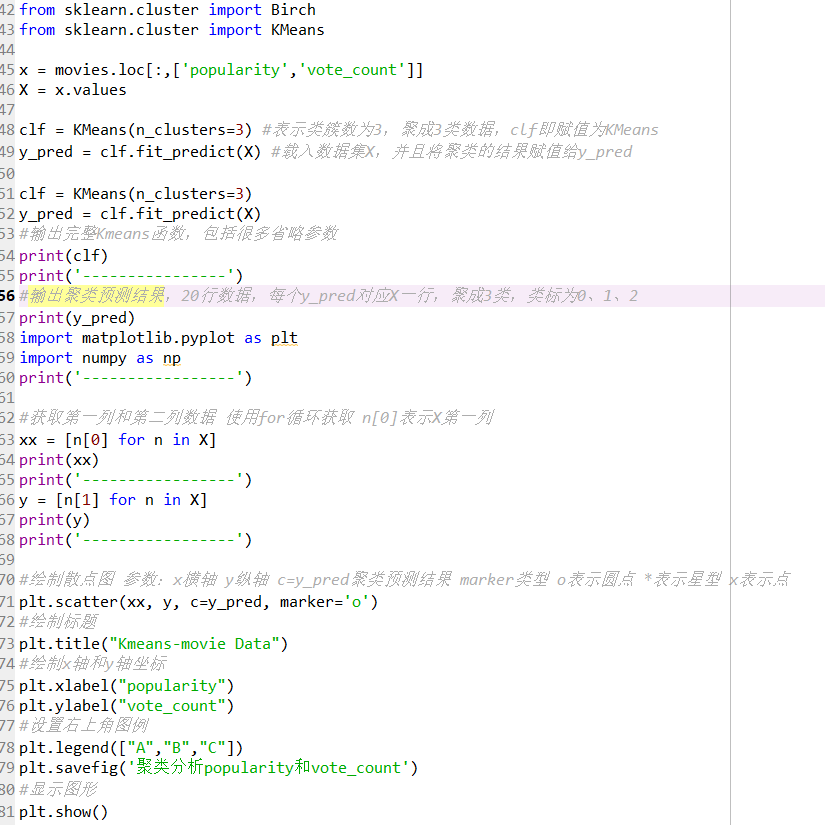
在TMDB上速度与激情8的评分是6.7分，可见预测模型的可行度还是比较高的。

问题六结论：经过预测模型预测的速度与激情8的评分为6.6，与TMDB上的6.7评分相差不多，可见预测模型可行度还是比较高的。

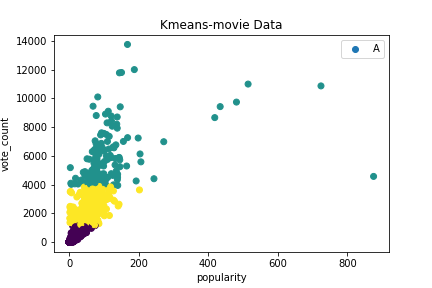
### 3.7问题七

问题7：如何使用Kmeans聚类对数据进行数据分析

解决方法：从数据集中挑选出‘popularity’和‘vote\_count’数据使用Kmeans聚类来进行分析。



**图3.7.1聚类分析代码图**



**图3.7.2聚类分析图**

问题七结论：经过使用K-Means聚类模型，并对分析结果可视化后可以看出，偏左下方的两部分数据的知名度（popularity）和评论次数（vote\_count）比较低，蓝色部分的知名度（popularity）和评论次数（vote\_count）则比较高，估计是影片当中比较火的类型，估计会深受大众的喜欢。

# 四结论和展望

## 4.1总结

首先在开头提出来的四个问题，经过我对数据的浏览、理解、数据清洗、建立模型和数据可视化后对解决了所提出来的四个问题并得出以下结论：

结论1：从上世纪90年代开始，整个电影市场呈现爆发式增长。其中，排名前五的戏剧类（Drama）、喜剧类（Comedy）、惊悚类（Thriller）、浪漫类（Romance）、动作类（Action）电影数量增长显著，排名前六至前九的类型增长相对比较缓慢。

结论2：（1）从图中可以看出不同类型的电影收益状况不同，其中又以动画类（Animation）和历史类（History）收益最高。

（2）年份与（Western）类型相关性最高，比其他类型的高出不少。

结论3： Universal Pictrues总票房收入高于Paramount Pictures，两公司排名前九的电影风格通过饼图展示出来，两公司的电影风格比例大致相同相差不多。

结论4：评价次数（Vote\_count）与票房收入最相关，电影预算（Budget）与票房收入高度相关，欢迎程度（Popularity）与评次高度相关，其他的因素与票房收入相关性相对较低。

结论5：各种电影类型的占比比例如图所示，其中占比类型最大的是Drama，占到了百分之18.1的比例，与第二名（comedy类型）拉开了百分之4.5的差距，可见电影类型中数量最多的就是Drama类型。

结论6：经过预测模型预测的速度与激情8的评分为6.6，与TMDB上的6.7评分相差不多，可见预测模型可行度还是比较高的。

结论7：经过使用K-Means聚类模型，并对分析结果可视化后可以看出，偏左下方的两部分数据的知名度（popularity）和评论次数（vote\_count）比较低，蓝色部分的知名度（popularity）和评论次数（vote\_count）则比较高，估计是影片当中比较火的类型，估计会深受大众的喜欢。

在分析数据过程中难的是对数据的理解，刚开始时对数据了解不够不能正确的处理数据，没有注意到数据中存在json格式，在这上面花了不少的时间。

## 4.2展望

首先我对数据的分析程度还不够，因为还可以通过数据来分析流行度（Popularity）给大众人群来推荐比较流行的电影，同时也可以通过电影的评分来给大众人群来推荐电影或者是按照国家分类来给群众推荐电影等，但是在分析的过程中没有意识到这些问题，是自己在分析数据的时候做的还不够好，希望在以后的学习中能够更全面的看待问题而不是仅仅只是片面的看待问题，希望能够更好的提升自己的能力。