**探索数据集**

1. 介绍

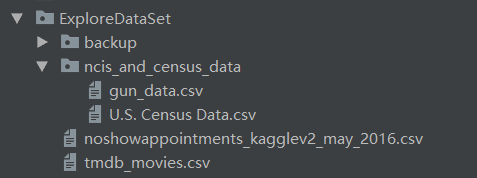
探索数据集程序分析了三个问题的数据，分别是：TMDb电影数据，未前往就诊的挂号预约，以及FBI枪支数据分析。可以在程序开始的时候选择需要分析哪一个数据集，随后程序会做出相关分析，提供统计数据结果及图形绘制。

TMDb电影数据的数据文件保存在tmdb\_movies.csv中；

未前往就诊的挂号预约数据文件保存在

noshowappointments\_kagglev2\_may\_2016.csv中；

FBI枪支数据文件保存在gun\_data.csv及U.S.Census Data.csv中。



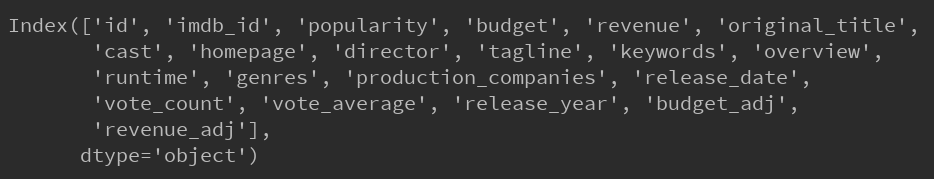
分析脚本文件保存为explore\_dataset.py文件。



1. 提出问题。

<1>.TMDb电影数据集分析

首先，通过df.columns查看TmDb数据集的数据类型，如下：

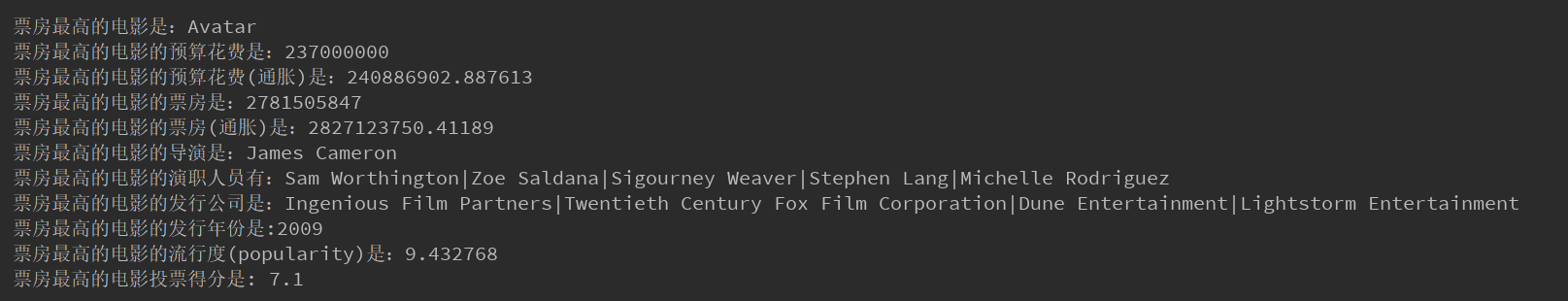


我们比较感兴趣的属性有：电影受欢迎程度(popularity)，电影预算(budget)，收入(revenue)，演职人员(cast)，电影类别(genres)，导演(director)，运行时间(runTime)，发行公司(production companies)，发布年份(release year)，平均投票得分(vote\_average)等等。基于以上属性，我们可以提出以下问题：

1. 票房最高的电影的演职人员是谁？导演是谁？发布公司是谁？是哪一年的？是什么电影等。
2. 哪位演职人员最近的一年票房最高？
3. 哪位演职人员的历史总票房最高？
4. 哪位导演最近一年的电影票房最高？
5. 哪位导演的历史总票房最高？
6. 哪家发行公司的电影历史总票房最高？
7. 最受欢迎（popularity）最高的电影类别(genres)是什么？
8. 平均投票得分(vote average)最高的电影类别(genres)是什么？
9. 票房高的电影有哪些特点？
10. 流行度最高的电影是什么？导演是谁？票房多少？
11. 评分最高的电影是什么？导演是谁？票房多少？
12. 超过1000人评价的电影中，平均评分最高的电影是什么？导演是谁？评分多少？
13. 数据处理及结果。

<1>.TMDb电影数据集分析

1. 票房最高的电影的演职人员是谁？导演是谁？发布公司是谁？是哪一年的？是什么电影等。



1. 哪位演职人员最近的一年票房最高？



1. 哪位演职人员的历史总票房最高？



1. 哪位导演最近一年的电影票房最高？



1. 哪位导演的历史总票房最高？



1. 哪家发行公司的电影历史总票房最高？



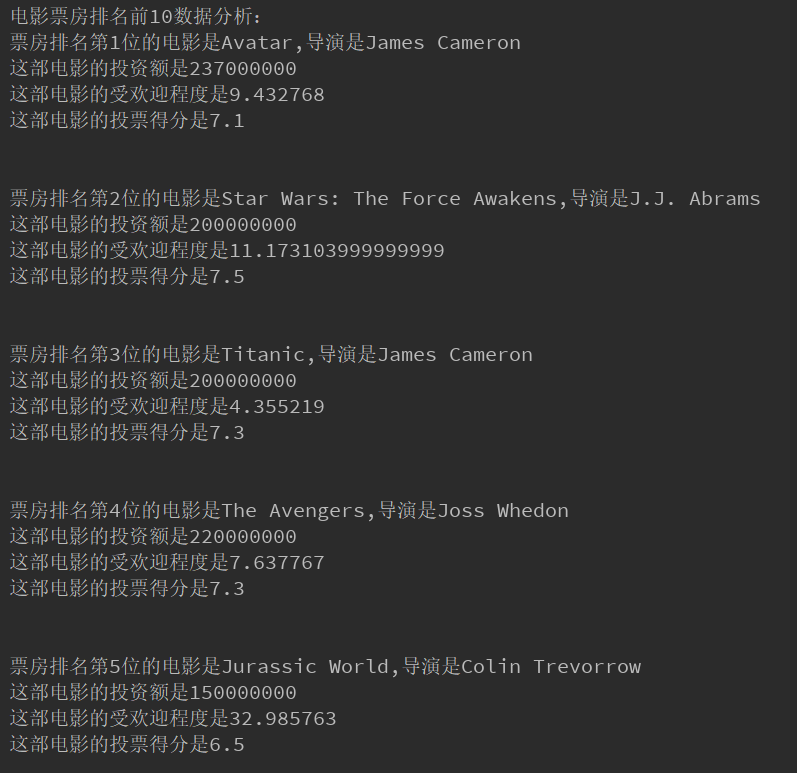
1. 最受欢迎（popularity）最高的电影类别(genres)是什么？

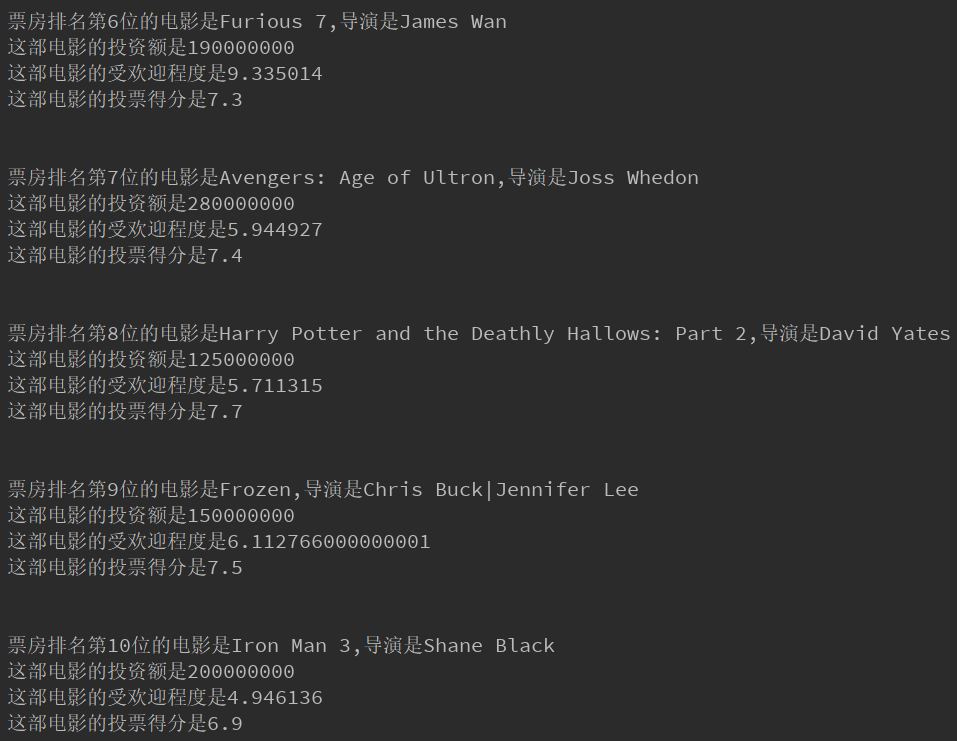


1. 平均投票得分(vote average)最高的电影类别(genres)是什么？



1. 票房高的电影有哪些特点？







1. 流行度最高的电影是什么？导演是谁？票房多少？



1. 评分最高的电影是什么？导演是谁？评分多少？



1. 超过1000人评价的电影中，平均评分最高的电影是什么？导演是谁？评分多少？

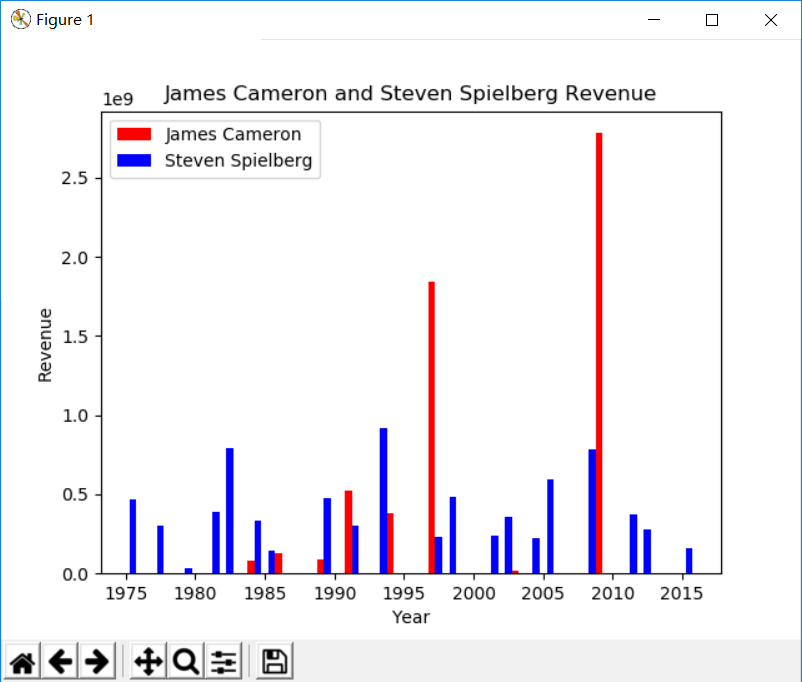


1. 图形绘制及统计结果分析。

图形绘制可以结合上面的问题，或提出新的问题进行绘制。

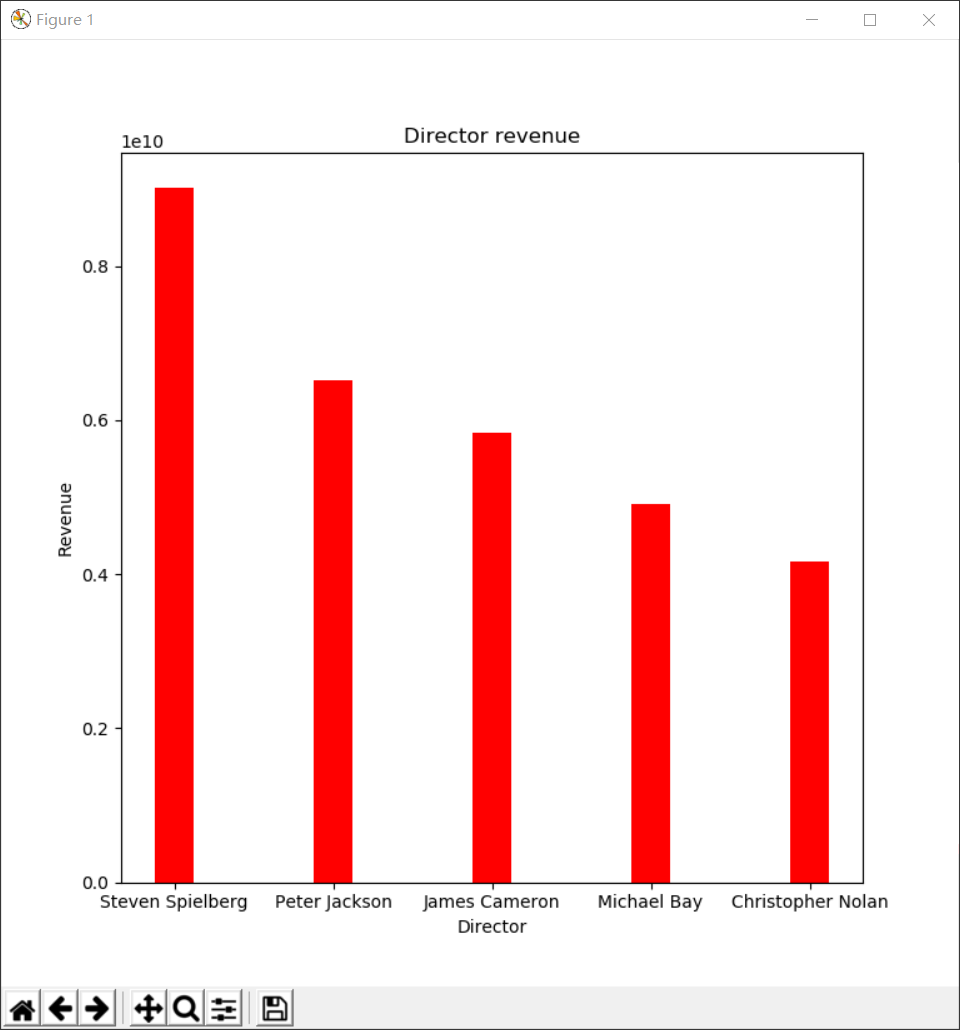
<1>.TMDb电影数据集分析

1. 从1975年到2015年，历年来詹姆斯卡梅隆（James Cameron）和斯蒂芬斯皮尔伯格(Steven Spielberg)票房对比。



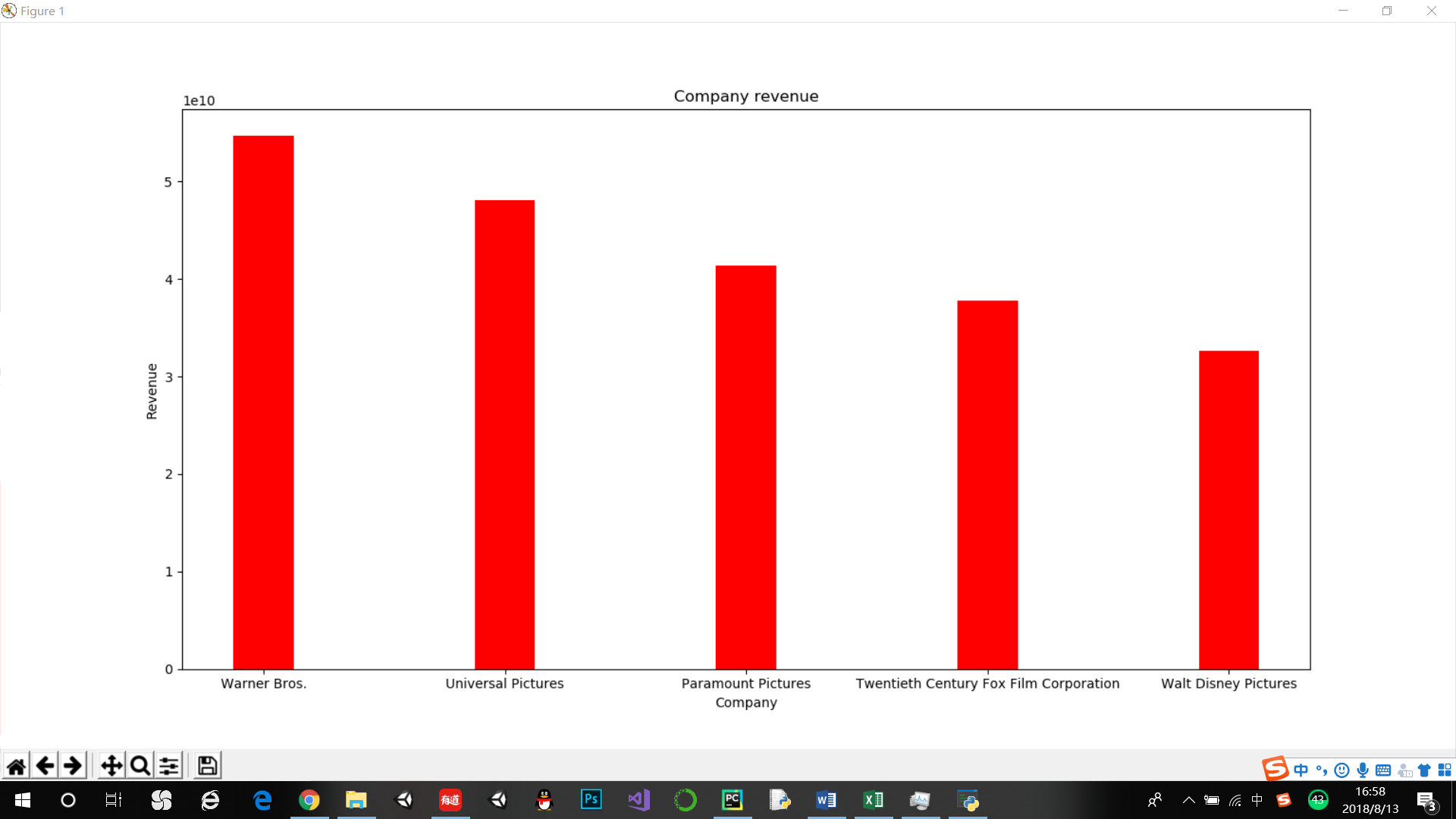
分析：从图片中可以看出，詹姆斯卡梅隆的作品数量比斯皮尔伯格要少很多，斯皮尔伯格从上个世纪70年代就有优秀的作品。同样作为世界级优秀电影导演，詹姆斯卡梅隆的作品虽然少，但是凭借泰坦尼克号和阿凡达（两根最高的红线），其在电影界的地位无人能撼动。

1. 显示历史票房排名前五的电影导演。



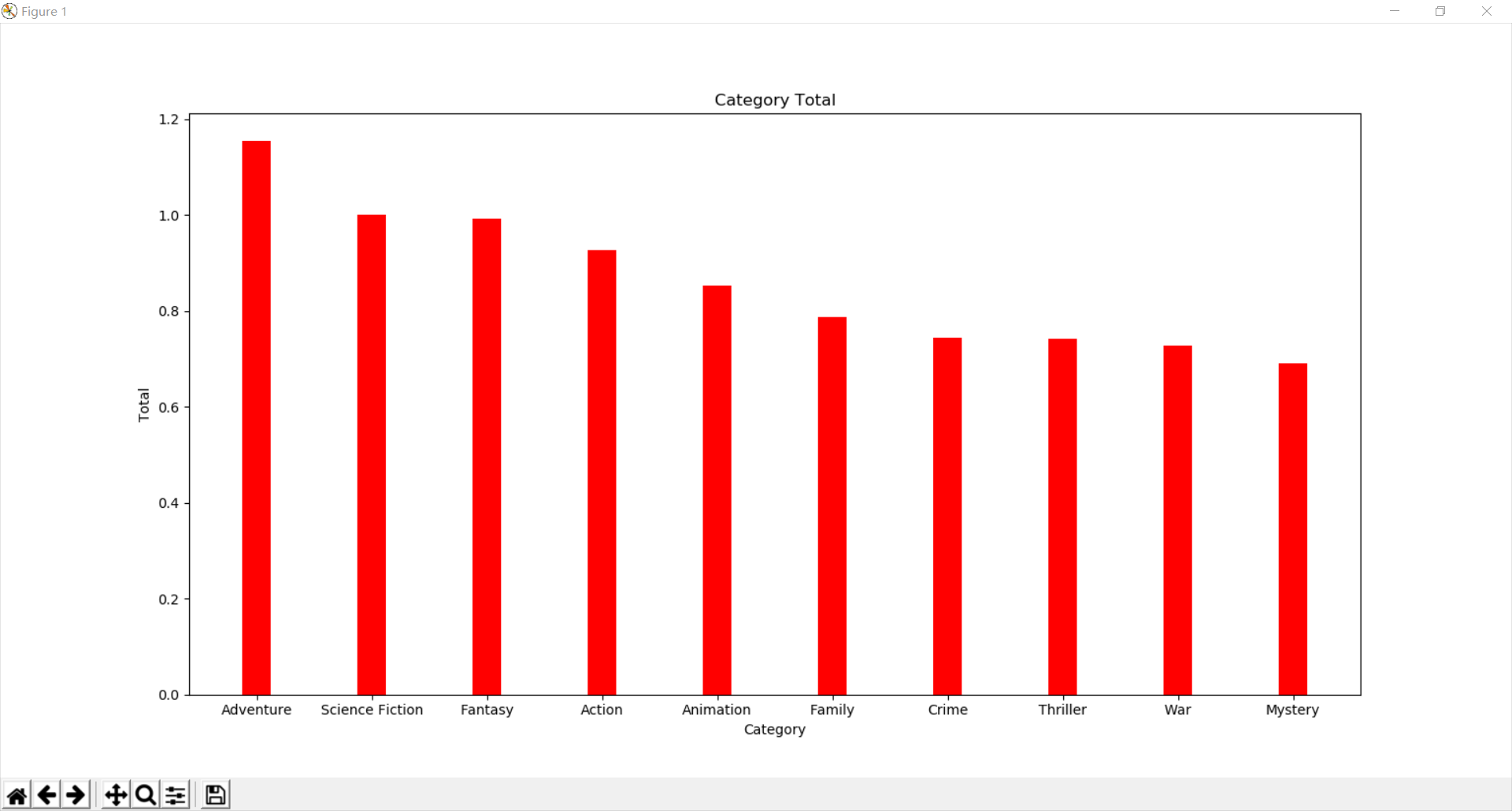
### 分析：历史票房前5的导演分别是斯蒂芬斯皮尔伯格，彼得杰克逊，詹姆斯卡梅隆，迈克尔·贝和[克里斯托弗·诺兰](http://www.baidu.com/link?url=ZPgkMSxZFzJIjDd9OYeXIIks0twlu3p8YxqBynGF9wSAq5ddGxYuoEkBWk2jPcOIgfdXmSkNIGHOOGqNx49U6a)。

1. 显示历史上获取票房最多的制片公司（前5）



分析：历史累计票房前五的发行公司分别是：华纳兄弟(Warner Bros)，环球影业(Universal Pictures)，派拉蒙影业公司(Paramount Pictures Company), 二十世纪福克斯公司(Tweentieth Century Fox)以及沃尔特迪士尼(Walt Disney)。

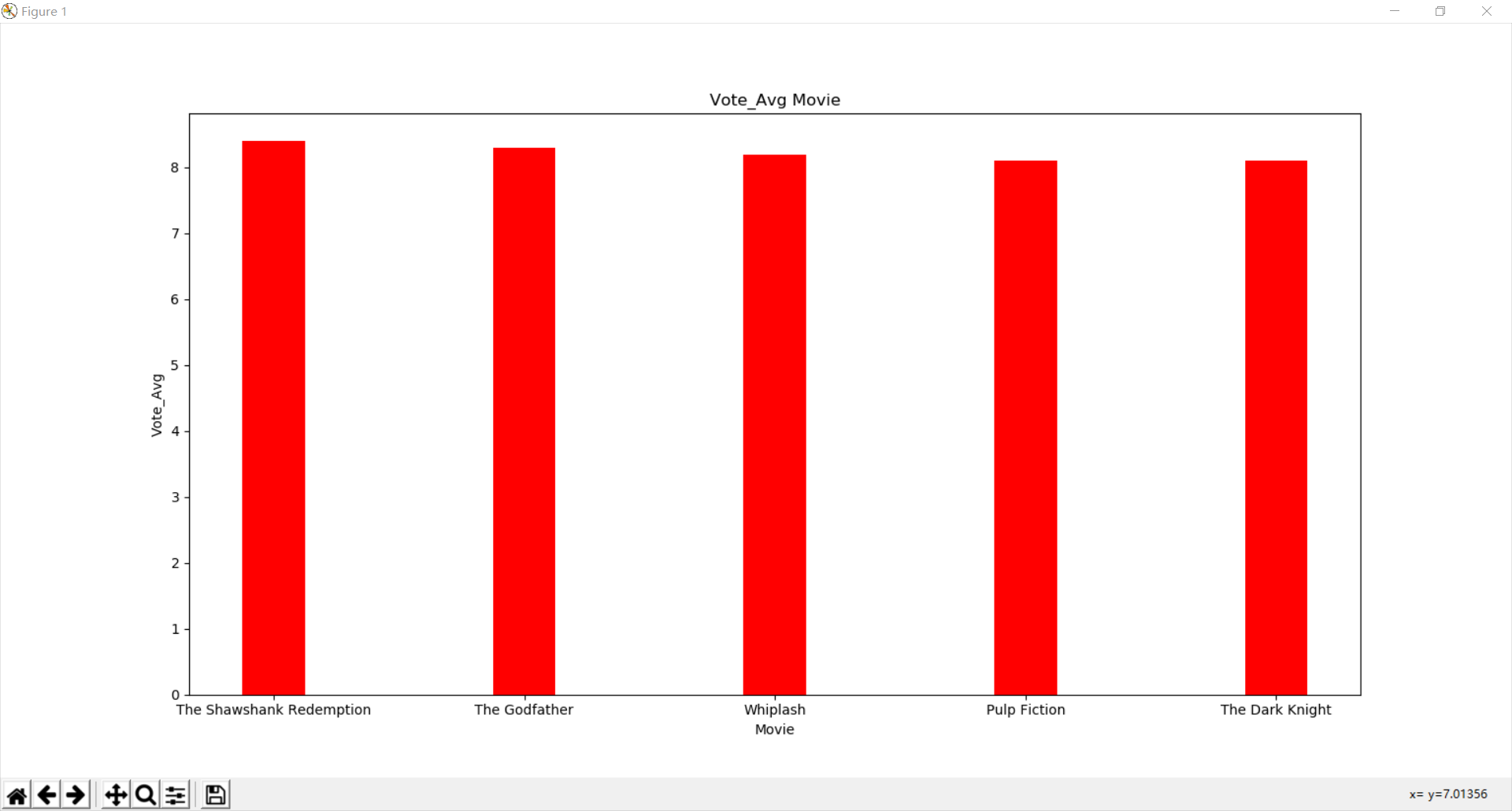
1. 最受欢迎的100部电影的类别统计。



分析：可以看出，最受欢迎的100部电影类别主要是：冒险类(Adventure),科幻类(Science Fiction), 奇幻类(Fantasy)，动作类(Action)，动画类(Animation)，

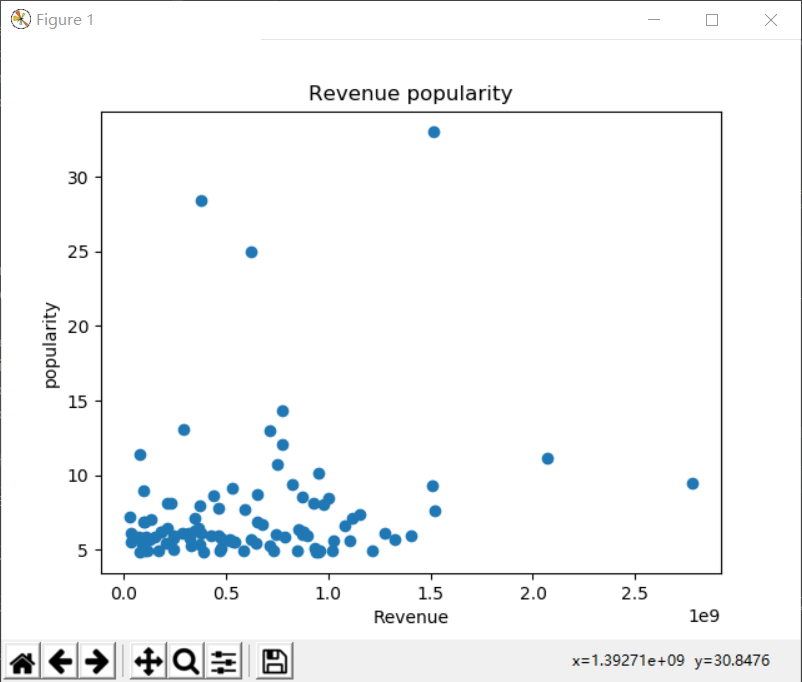
家庭类(Family),犯罪类(Crime),惊悚类(Thriller),战争类(War),推理类(Mystery)等。

1. 超过1000人的电影评价中，评价前五名电影是什么。



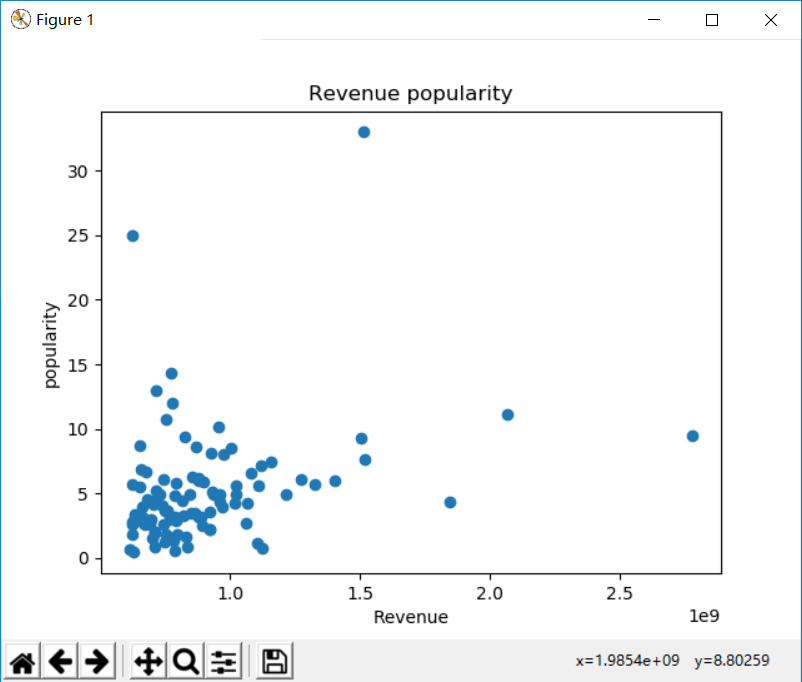
分析：超过1000人的电影评价中，评价前五名电影分别是肖申克的救赎,教父，爆裂鼓手,低俗小说以及黑暗骑士。

6.在流行度排名前100的电影中，绘制流行度与票房收入的散点图，二者是否有关联。

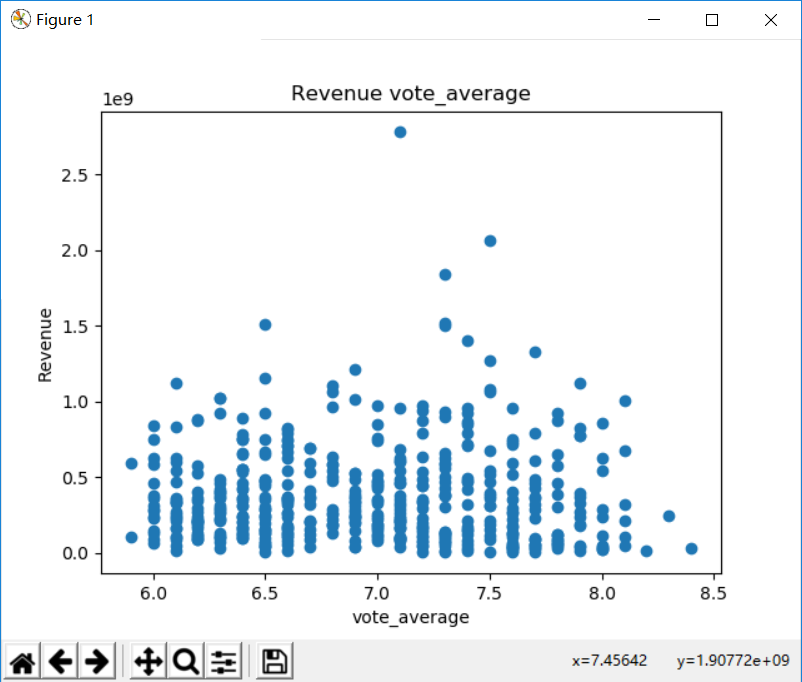


分析：我认为流行度排名前100的电影其票房收入并没有和票房收入成正向相关分布。

7.票房排名前100名的电影，其流行度与票房的散点图。

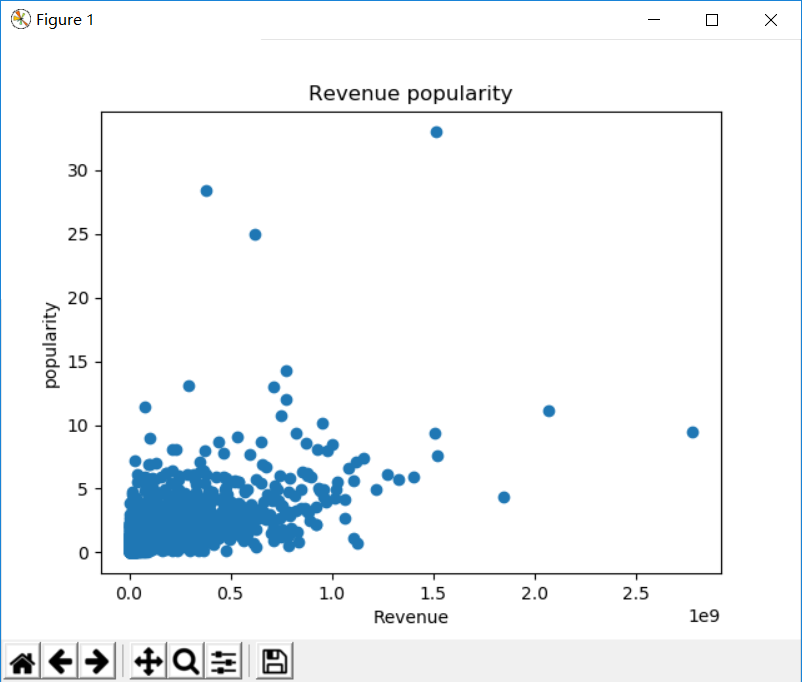


8.1000人以上的评分的电影中，评分前500名的电影评分与电影票房的关系。

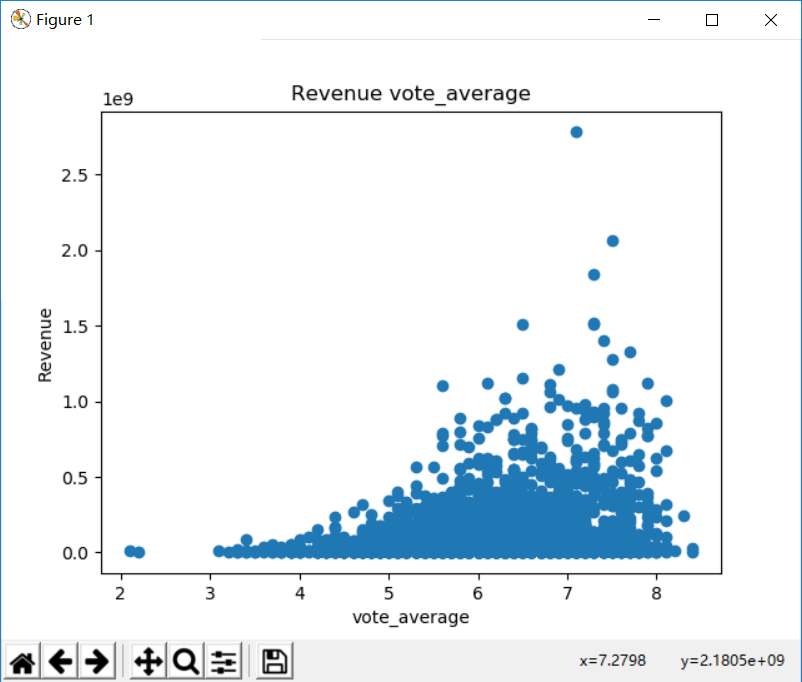


分析：从图中可以看出，从6-8分都有票房高的电影和票房低的电影。没有显著的差异。

9.查看所有电影流行度与票房间的关系。

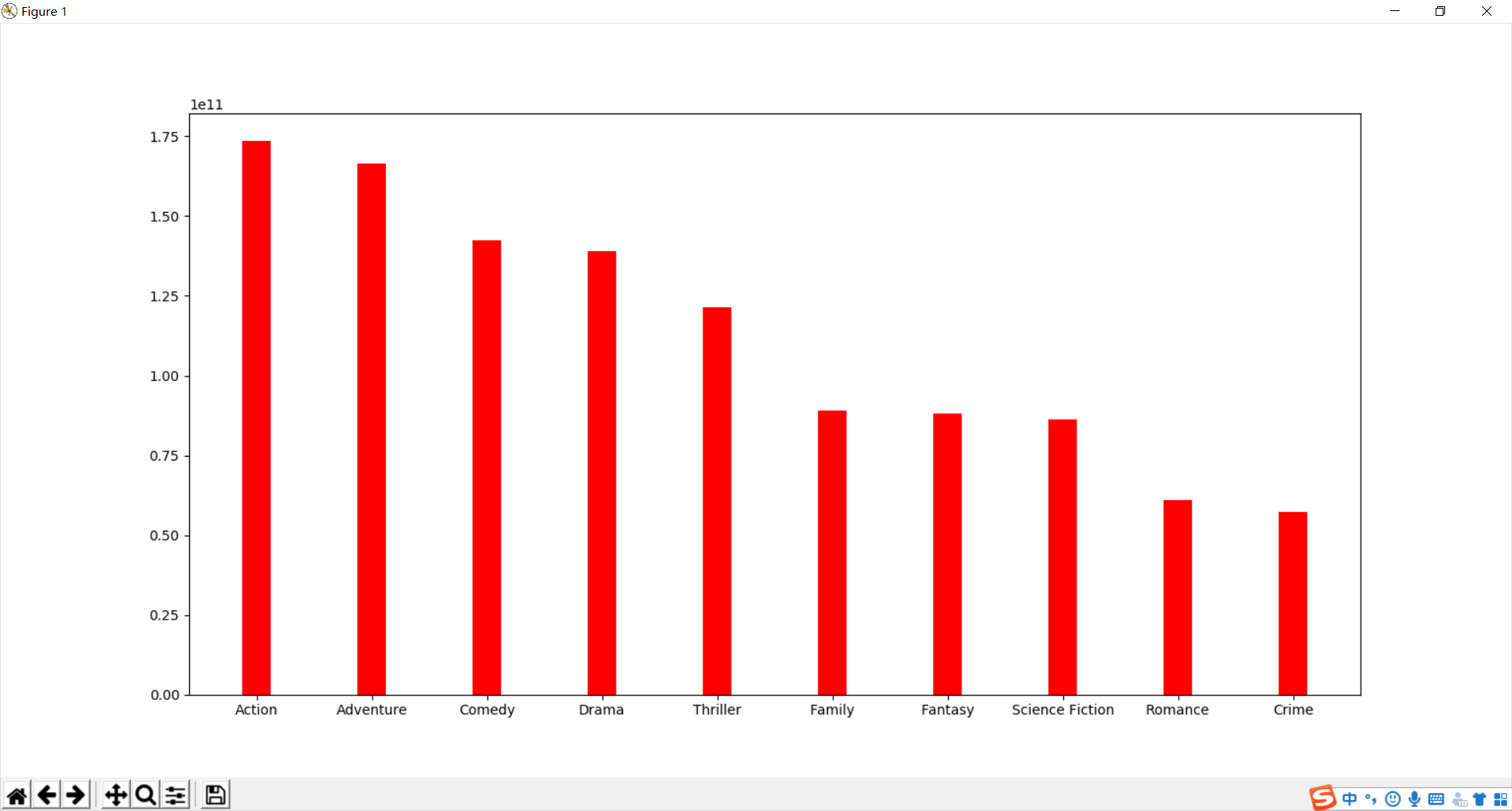


10.查看所有评分与电影票房间的关系。



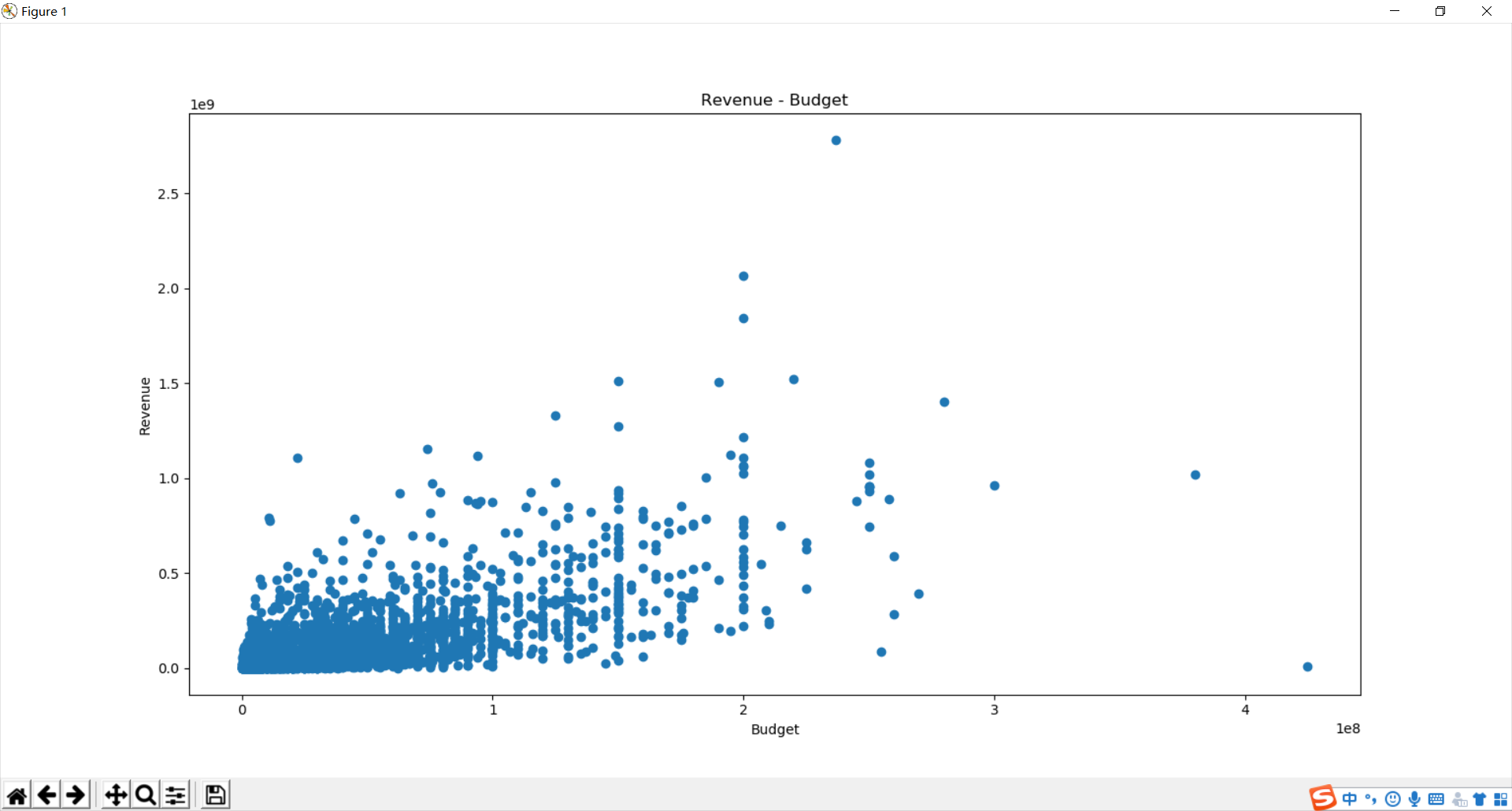
分析：从这张图可以看出，高票房电影的评分大多集中在6-8分间。超高票房的电影收入一般大于6分。评分小与5分的电影几乎没有高票房。

11.查看不同类型的电影的票房对比。

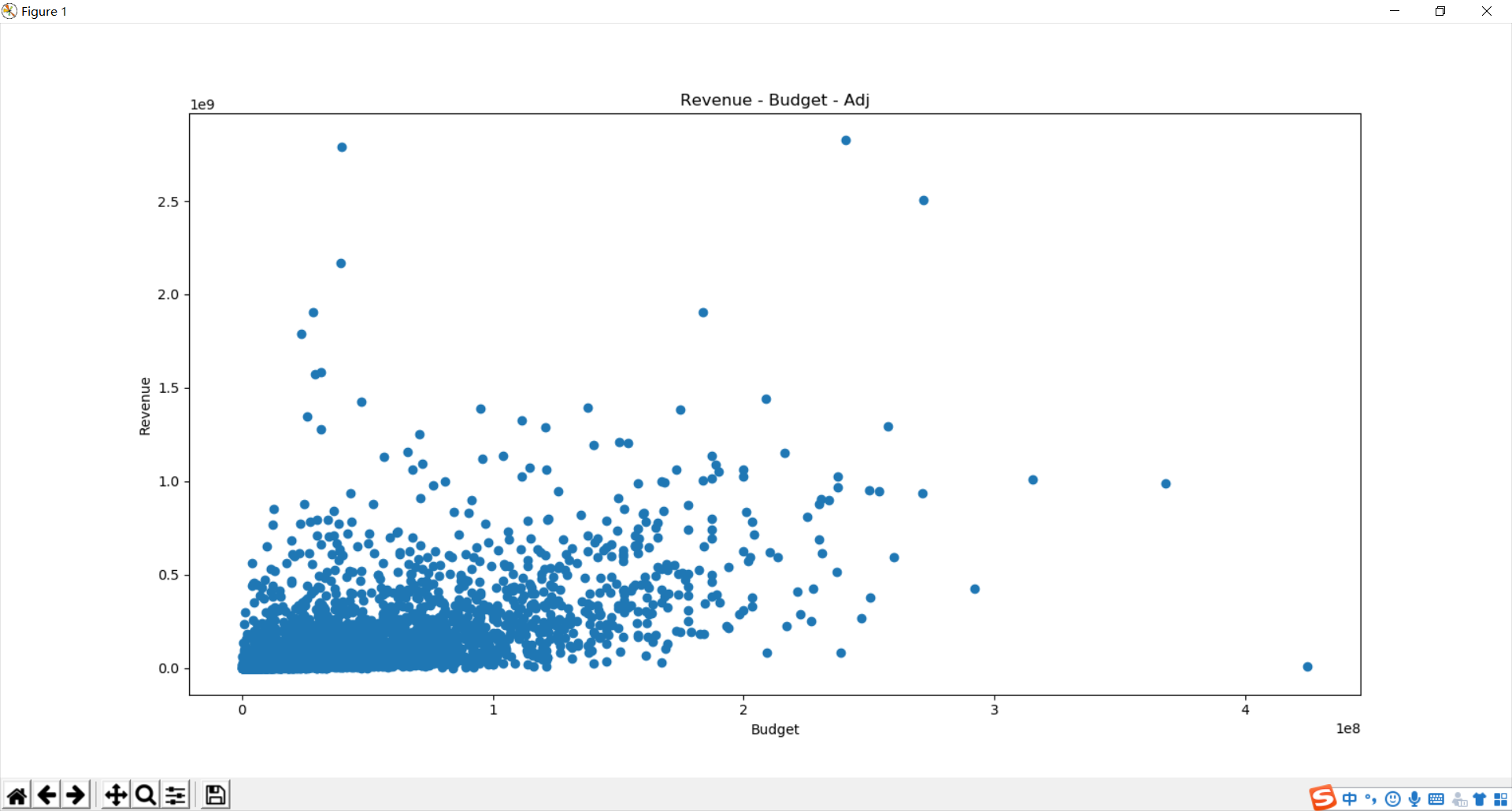


分析：动作类，冒险类，喜剧类，戏剧类，惊悚类电影的票房要高于其它类型的电影。

12.电影预算与电影票房间的关系。



分析：可以看出，随着预算的增加，票房有了一定的增加，但不明显。低预算也有票房不低的电影，高预算也有低票房的电影。但是要想得到极高的票房（进入票房排名），则一定的预算是必要的。小成本的电影虽然能取得好的票房，但很难取得巨额的票房。



但是考虑通胀情况，低成本投资的电影里面也有很多高收入的电影。

1. 总结。

通过数据处理及图形绘制和统计结果分析，可以得出一些结论：

<1>.TMDb电影数据集分析

1. 史上最高票房的电影是阿凡达，导演是詹姆斯卡梅隆；
2. 史上累计票房最高的电影导演是斯皮尔伯格；
3. 史上累计票房最高的电影演员是哈里森福特；
4. 侏罗纪世界是目前流行度最高的电影;
5. 评分超过1000人的电影中，肖申克的救赎评分最高;
6. 华纳兄弟公司是史上票房最多的电影公司;
7. 最受欢迎的100部作品中，冒险类(Adventure),科幻类(Science Fiction), 奇幻类(Fantasy)，动作类(Action)，动画类(Animation)这些类型的比例最高；
8. 高票房的电影评分大多集中在6分以上（6-8分）；
9. 动作类，冒险类，喜剧类，戏剧类，惊悚类电影的票房要高于其它类型的电影；
10. 不考虑通胀的情况下，票房和预算不完全成正相关，但是好的票房（非常高的票房）需要一定的预算支出；
11. 开发环境

Win10下 PyCharm开发。

