探索数据集

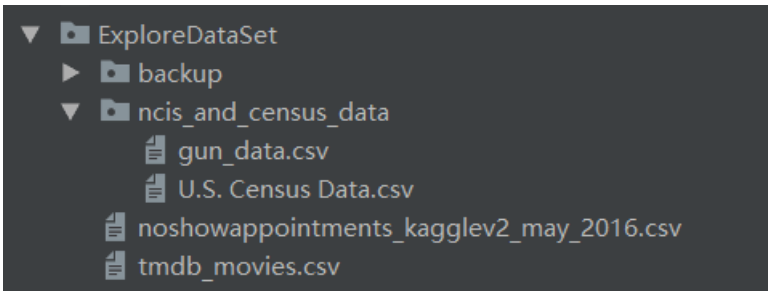
一．介绍

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。

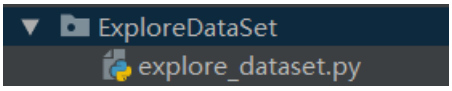
TMDb 电影数据的数据文件保存在 tmdb_movies.csv 中；

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。



二．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：

```
Index(['id', 'imdb_id', 'popularity', 'budget', 'revenue', 'original_title',
       'cast', 'homepage', 'director', 'tagline', 'keywords', 'overview',
       'runtime', 'genres', 'production_companies', 'release_date',
       'vote_count', 'vote_average', 'release_year', 'budget_adj',
       'revenue_adj'],
      dtype='object')
```

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?

三．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。

```
票房最高的电影是：Avatar
票房最高的电影的预算花费是：237000000
票房最高的电影的预算花费(通胀)是：240886902.887613
票房最高的电影的票房是：2781505847
票房最高的电影的票房(通胀)是：2827123750.41189
票房最高的电影的导演是：James Cameron
票房最高的电影的演职人员有：Sam Worthington|Zoe Saldana|Sigourney Weaver|Stephen Lang|Michelle Rodriguez
票房最高的电影的发行公司是：Ingenious Film Partners|Twentieth Century Fox Film Corporation|Dune Entertainment|Lightstorm Entertainment
票房最高的电影的发行年份是:2009
票房最高的电影的流行度(popularity)是：9.432768
票房最高的电影投票得分是：7.1
```

2. 哪位演职人员最近的一年票房最高?

```
最近1年(2015)年演员Harrison Ford的票房总收入最高，总票房为：2110808001
```

3. 哪位演职人员的历史总票房最高?

```
演员Harrison Ford的历史总票房最高，总票房为：8922840695
```

4. 哪位导演最近一年的电影票房最高?

```
最近1年(2015)年导演J.J. Abrams的票房总收入最高，总票房为：2068178225
```

5. 哪位导演的历史总票房最高?

```
导演Steven Spielberg的历史总票房最高，总票房为：9018563772
```

6. 哪家发行公司的电影历史总票房最高?

```
Warner Bros.公司的历史总票房最高，总票房为：54688433698
```

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

```
Adventure类型的电影受欢迎程度最高，平均受欢迎程度为1.1542590441876264
```

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

```
Documentary类型的电影平均投票得分最高，平均投票得分为6.908461538461542
```

9. 票房高的电影有哪些特点?

```
电影票房排名前10数据分析:
票房排名第1位的电影是Avatar,导演是James Cameron
这部电影的投资额是237000000
这部电影的受欢迎程度是9.432768
这部电影的投票得分是7.1


票房排名第2位的电影是Star Wars: The Force Awakens,导演是J.J. Abrams
这部电影的投资额是200000000
这部电影的受欢迎程度是11.173103999999999
这部电影的投票得分是7.5


票房排名第3位的电影是Titanic,导演是James Cameron
这部电影的投资额是200000000
这部电影的受欢迎程度是4.355219
这部电影的投票得分是7.3


票房排名第4位的电影是The Avengers,导演是Joss Whedon
这部电影的投资额是220000000
这部电影的受欢迎程度是7.637767
这部电影的投票得分是7.3


票房排名第5位的电影是Jurassic World,导演是Colin Trevorrow
这部电影的投资额是150000000
这部电影的受欢迎程度是32.985763
这部电影的投票得分是6.5
```

```
票房排名第6位的电影是Furious 7,导演是James Wan
这部电影的投资额是190000000
这部电影的受欢迎程度是9.335014
这部电影的投票得分是7.3


票房排名第7位的电影是Avengers: Age of Ultron,导演是Joss Whedon
这部电影的投资额是280000000
这部电影的受欢迎程度是5.944927
这部电影的投票得分是7.4


票房排名第8位的电影是Harry Potter and the Deathly Hallows: Part 2,导演是David Yates
这部电影的投资额是125000000
这部电影的受欢迎程度是5.711315
这部电影的投票得分是7.7


票房排名第9位的电影是Frozen,导演是Chris Buck|Jennifer Lee
这部电影的投资额是150000000
这部电影的受欢迎程度是6.112766000000001
这部电影的投票得分是7.5


票房排名第10位的电影是Iron Man 3,导演是Shane Black
这部电影的投资额是200000000
这部电影的受欢迎程度是4.946136
这部电影的投票得分是6.9
```

```
电影票房排名前10的电影里面，演员出演的次数：
actor
Robert Downey Jr.       3
Chris Evans             2
Scarlett Johansson      2
Michelle Rodriguez      2
Mark Ruffalo            2
Chris Hemsworth         2
Eva Marie Saint         1
Frances Fisher          1
Zoe Saldana             1
Emma Bell               1
George C. Scott         1
Guy Pearce              1
Gwyneth Paltrow         1
Harrison Ford           1
```

```
电影票房排名前10的电影里面，电影发行公司有：
company
Marvel Studios                          3
Twentieth Century Fox Film Corporation  2
Dentsu                                  2
Lightstorm Entertainment                2
Warner Bros.                            2
Universal Pictures                      1
Truenorth Productions                   1
Amblin Entertainment                    1
American Zoetrope                       1
ArieScope Pictures                      1
Bad Robot                               1
Walt Disney Animation Studios           1
Dune Entertainment                      1
Fuji Television Network                 1
Heyday Films                            1
Ingenious Film Partners                 1
Jerry Weintraub Productions             1
```

```
电影票房排名前10的电影里面，电影类型为：
category
Adventure          8
Action             8
Science Fiction    6
Thriller           5
Fantasy            3
Romance            2
Family             2
Drama              2
Crime              1
Animation          1
```

10. 流行度最高的电影是什么？导演是谁？票房多少？

```
流行度最高的电影是Jurassic World,导演是Colin Trevorrow,票房是1513528810
```

11. 评分最高的电影是什么？导演是谁？评分多少？

```
平均评分最高的电影是The Story of Film: An Odyssey,导演是Mark Cousins, 平均得分是9.2
```

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？

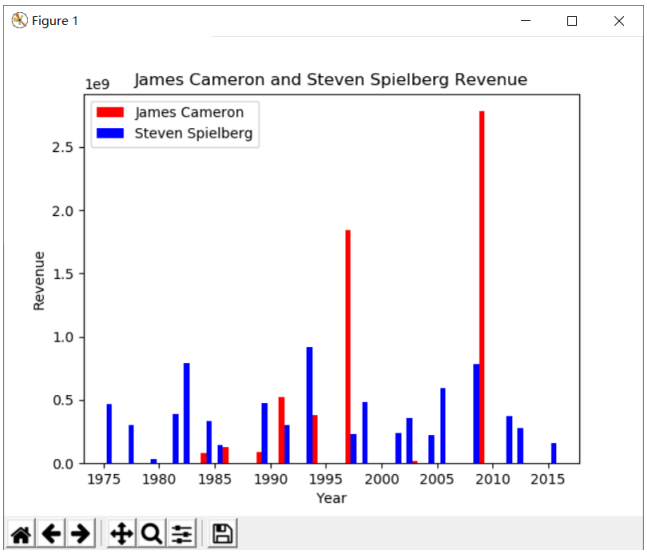
```
超过1000人评价的电影中，平均评分最高的电影是The Shawshank Redemption,导演是Frank Darabont, 平均得分是8.4
```

四．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。

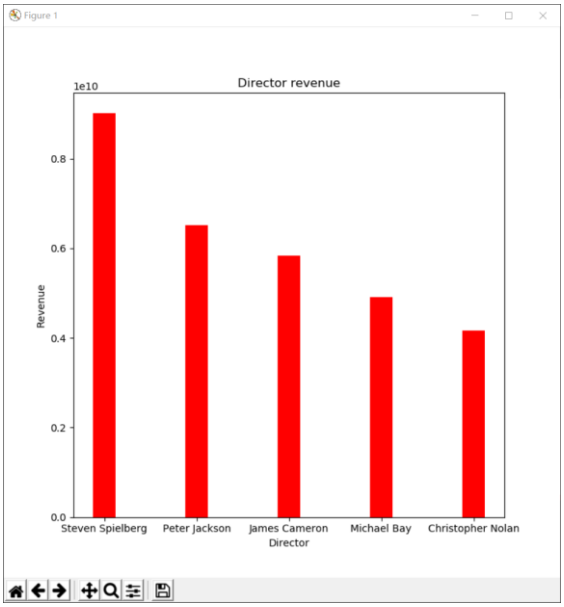
<1>.TMDb 电影数据集分析

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。



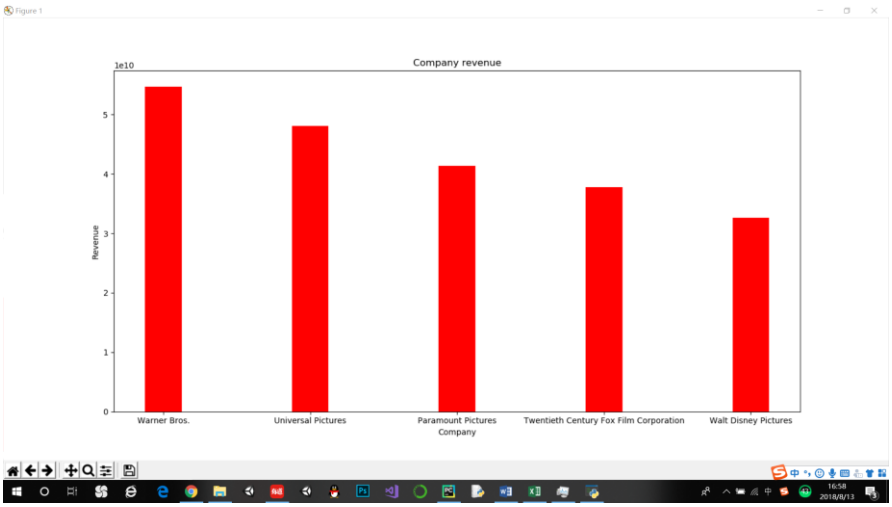
分析：从图片中可以看出，詹姆斯卡梅隆的作品数量比斯皮尔伯格要少很多，斯皮尔伯格从上个世纪 70 年代就有优秀的作品。同样作为世界级优秀电影导演，詹姆斯卡梅隆的作品虽然少，但是凭借泰坦尼克号和阿凡达（两根最高的红线），其在电影界的地位无人能撼动。

2. 显示历史票房排名前五的电影导演。



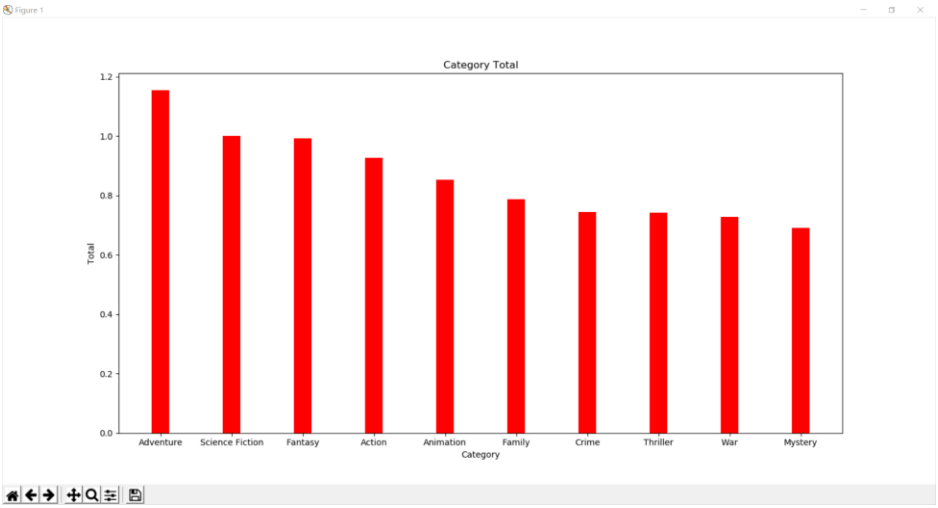
分析：历史票房前 5 的导演分别是斯蒂芬斯皮尔伯格，彼得杰克逊，詹姆斯卡梅隆，迈克尔·贝和克里斯托弗·诺兰。

3. 显示历史上获取票房最多的制片公司（前 5）



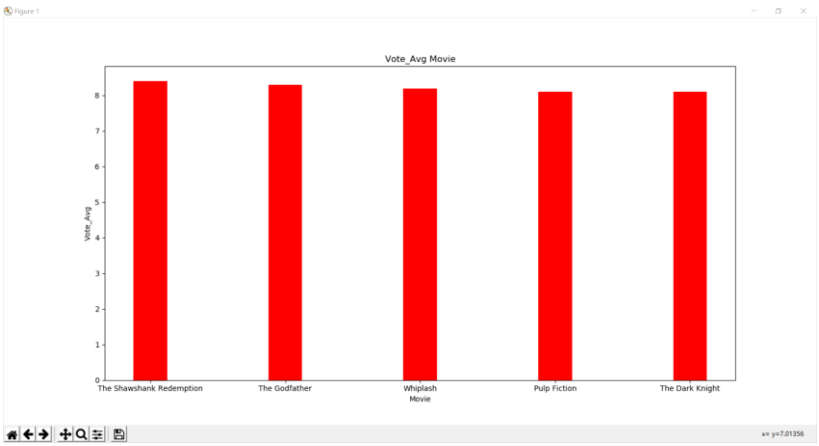
分析：历史累计票房前五的发行公司分别是：华纳兄弟(Warner Bros)，环球影业(Universal Pictures)，派拉蒙影业公司(Paramount Pictures Company)，二十世纪福克斯公司(Tweentieth Century Fox)以及沃尔特迪士尼(Walt Disney)。

4. 最受欢迎的 100 部电影的类别统计。



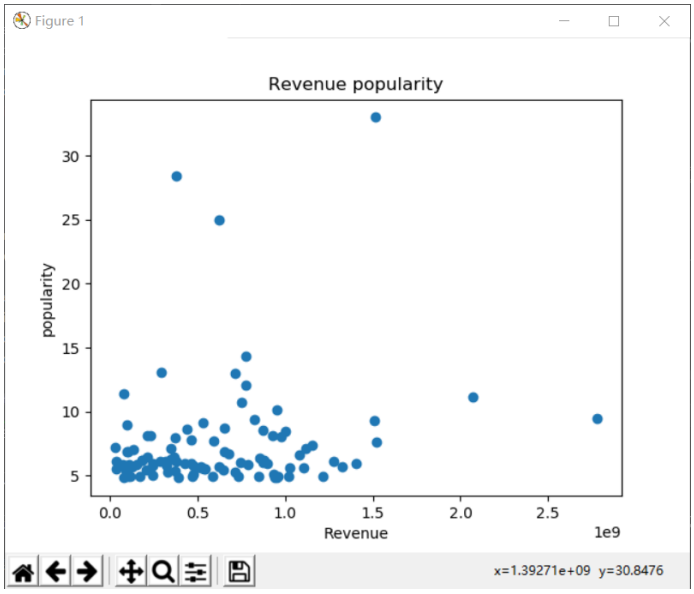
分析：可以看出，最受欢迎的 100 部电影类别主要是：冒险类(Adventure)，科幻类(Science Fiction)，奇幻类(Fantasy)，动作类(Action)，动画类(Animation)，家庭类(Family),犯罪类(Crime),惊悚类(Thriller),战争类(War),推理类(Mystery)等。

5. 超过 1000 人的电影评价中，评价前五名电影是什么。



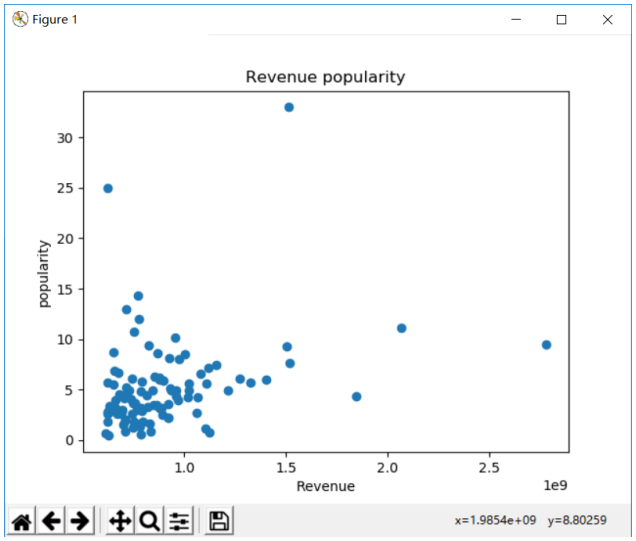
分析：超过 1000 人的电影评价中，评价前五名电影分别是肖申克的救赎,教父，爆裂鼓手,低俗小说以及黑暗骑士。

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。

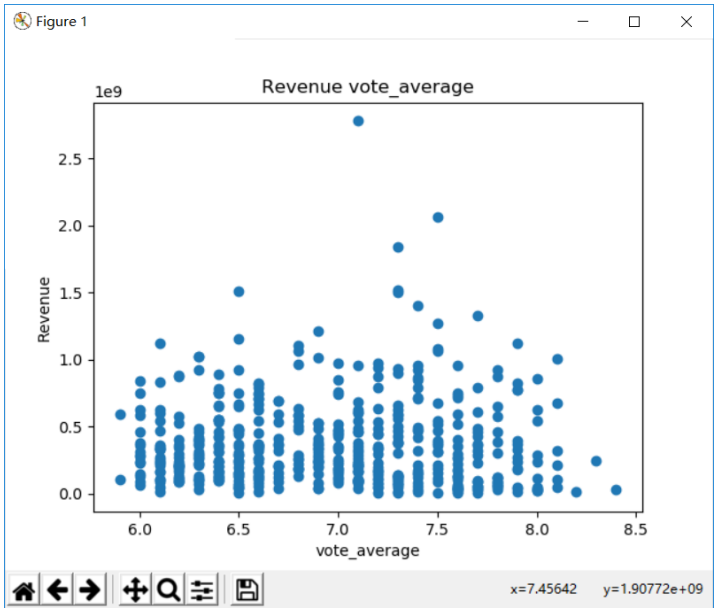


分析：我认为流行度排名前 100 的电影其票房收入并没有和票房收入成正向相关分布。

7.票房排名前 100 名的电影，其流行度与票房的散点图。

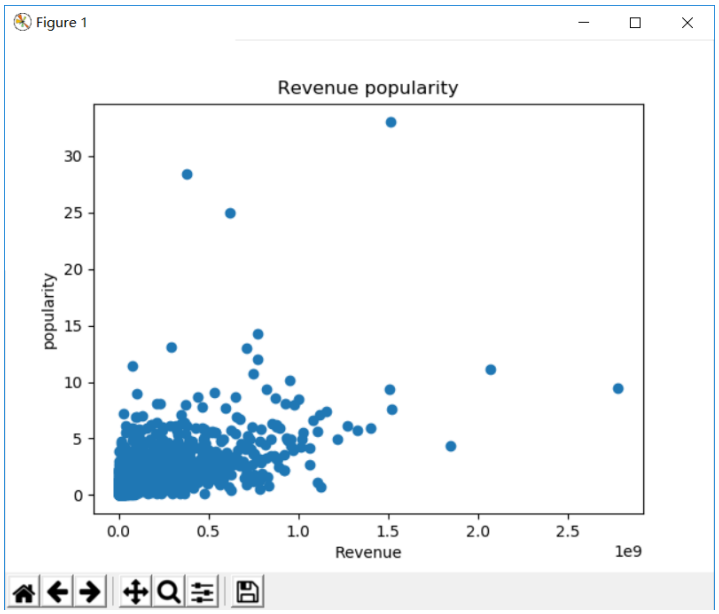


8.1000 人以上的评分的电影中，评分前 500 名的电影评分与电影票房的关系。

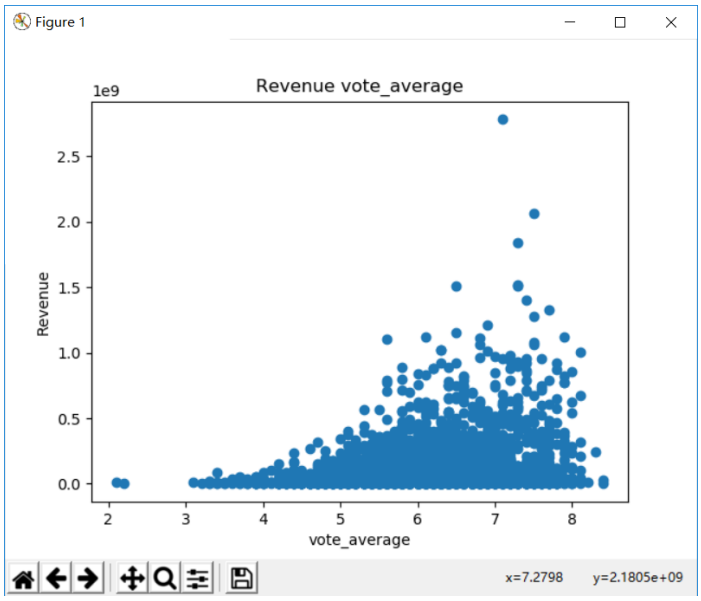


分析：从图中可以看出，从 6-8 分都有票房高的电影和票房低的电影。没有显著的差异。

9.查看所有电影流行度与票房间的关系。

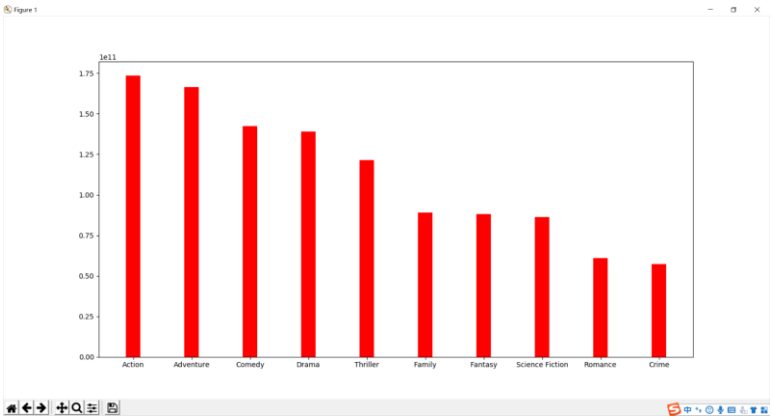


10.查看所有评分与电影票房间的关系。



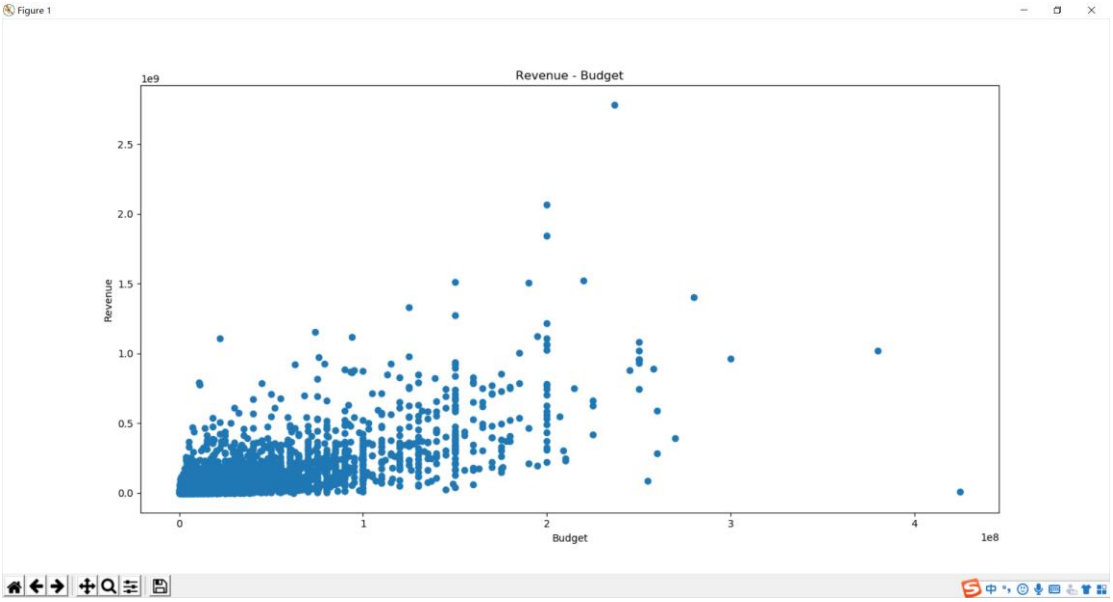
分析：从这张图可以看出，高票房电影的评分大多集中在 6-8 分间。超高票房的电影收入一般大于 6 分。评分小与 5 分的电影几乎没有高票房。

11.查看不同类型的电影的票房对比。

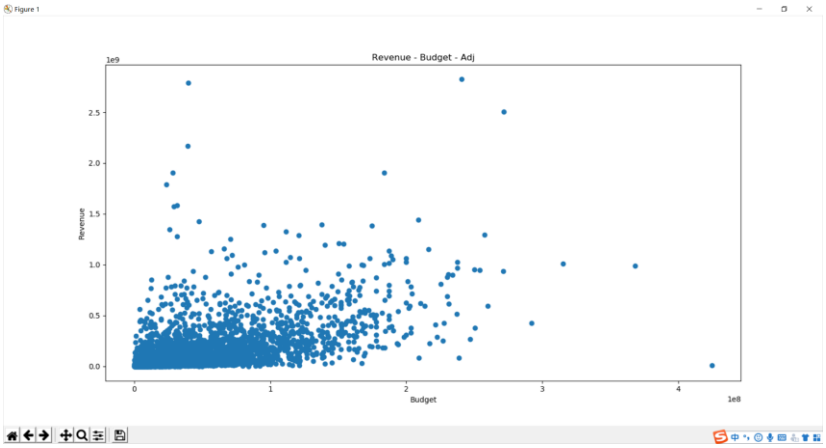


分析：动作类，冒险类，喜剧类，戏剧类，惊悚类电影的票房要高于其它类型的电影。

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。

五．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；

六．开发环境

Win10 下 PyCharm 开发。