开放性数据分析--TMDb 电影数据

1 数据描述

1.1 选择的数据内容

选择的数据为 TMDb 电影数据

1.2 数据特征说明

该数据共 21 列特征，解释如下：

列序号	列名	列解释	备注
0	id	电影 ID	
1	imdb_id	Imdb_id	
2	popularity	欢迎程度	百分比
3	budget	预算	美元
4	revenue	电影收入	美元
5	original_title	电影名称	
6	cast	演员表	
7	homepage	电影网址	
8	director	导演	
9	tagline	宣传词	
10	keywords	关键词	
11	overview	剧情摘要	
12	runtime	电影时长	
13	genres	电影风格	
14	production_companies	制作公司	
15	release_date	发布日期	
16	vote_count	评价次数	
17	vote_average	平均评分	
18	release_year	发布年份	
19	budget_adj	预算(考虑通货膨胀因素调整)	
20	revenue_adj	电影收入(考虑通货膨胀因素调整)	

2 提出的问题及分析结果

以下排除评价次数小于 50 的电影（样本太小，评分不可信）
排除预算和电影收入小于 100 美元的电影（明显不符合常理）.
分析过程中，如果没有特殊说明，使用的”预算”和”收入”均使用考虑通货膨胀因素调整后的值。

2.1 数据检查

首先按上述条件排除数据；
而后使用 info 函数检查数据，如下：

id	3126 non-null int64
imdb_id	3126 non-null object
popularity	3126 non-null float64
budget	3126 non-null int64
revenue	3126 non-null int64
original_title	3126 non-null object
cast	3126 non-null object
homepage	1276 non-null object
director	3126 non-null object
tagline	2987 non-null object
keywords	3063 non-null object
overview	3126 non-null object
runtime	3126 non-null int64
genres	3126 non-null object
production_companies	3122 non-null object
release_date	3126 non-null object
vote_count	3126 non-null int64
vote_average	3126 non-null float64
release_year	3126 non-null int64
budget_adj	3126 non-null float64
revenue_adj	3126 non-null float64

在这个数据中，出现 NAN 的数据列包括：homepage/tagline/keywords/production_companies。这些数据都不是我们关注的，故不作进一步处理。

2.2 收入高的电影有哪些特点

电影的票房是投资商们普遍关注的问题。我们预期一个电影的收入可能与以

下因素有关：

1. 电影预算：投入越高的大制作越能获得更多的票房；
2. 电影风格：某些风格的电影会更受观众们的欢迎；
3. 电影评分：电影评分代表电影的艺术水平,高水平的电影有可能得到更多的票房;
4. 电影受欢迎程度:人气较高的电影应该会有更多的票房

2.2.1 将电影按照收入分级

对”通货膨胀因素调整后的电影收入”一项进行统计,

根据统计规律，所有电影中票房最高的是阿凡达， 28.27 亿美元，收入最低的只有 155 美元。其它统计数字：

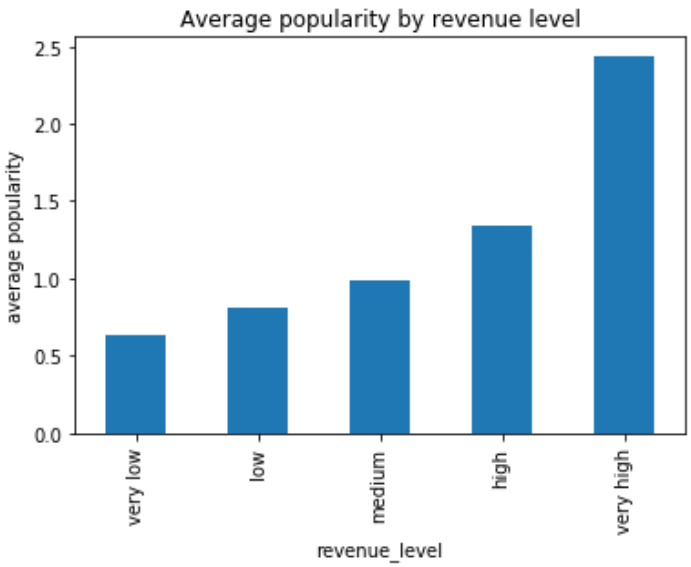
统计内容	值（美元）
最低	155
25%	316 万
50%	828 万
75%	1.96 亿
最高	28.27 亿

根据这一统计数字，我们将电影收入分为 5 个等级，并统计各个等级的电影数字如下：

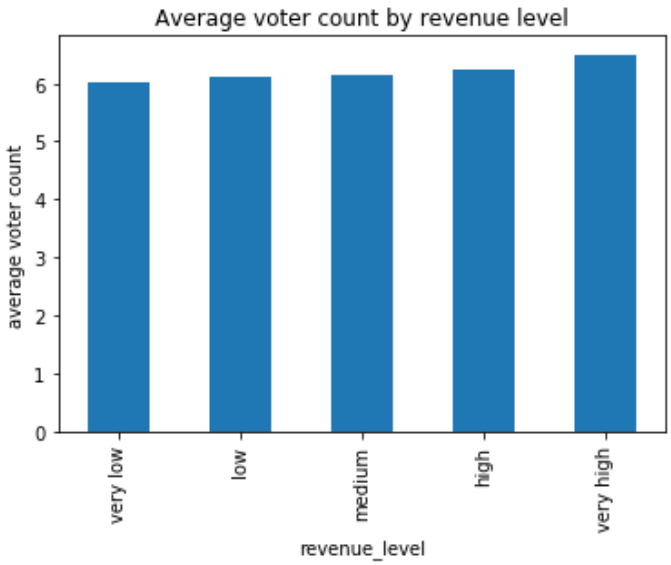
等级	预算	电影数量
很低	<10 万美元	87
低	10 万-316 万美元	694
中	316 万-828 万美元	781
高	828 万-1.96 亿美元	781
很高	>1.96 亿美元	781

2.2.2 比较不同收入等级的各项电影特征

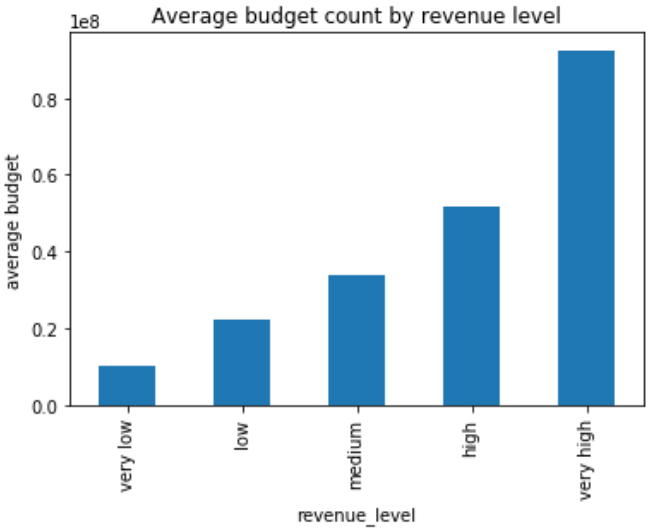
各收入等级电影平均受欢迎程度：



各收入等级电影平均评分：



各收入等级电影平均预算：



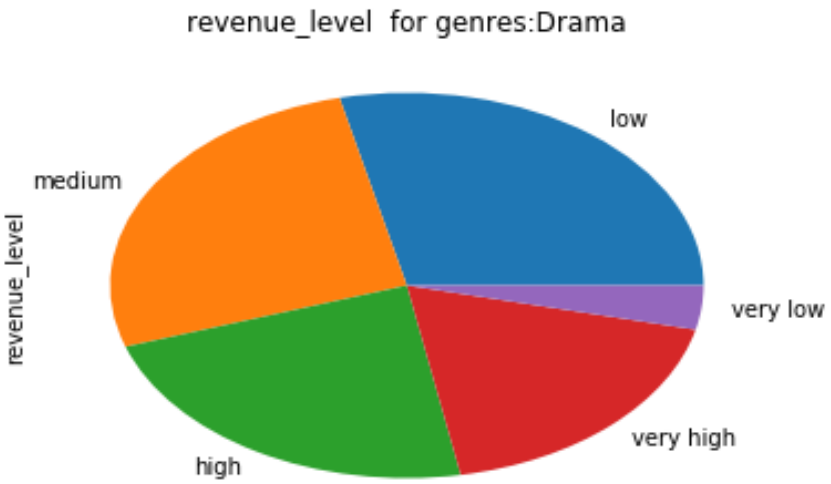
2.2.3 比较各个风格的电影中，不同收入等级所占比例

每个电影通常含有 1 个或若干个风格，为了统计不同收入等级中电影风格的比例，对原数据进行修改，将包含多个电影风格的电影拆成多行，每行仅包含 1 个电影风格。而后绘制各个风格的电影收入等级对比。

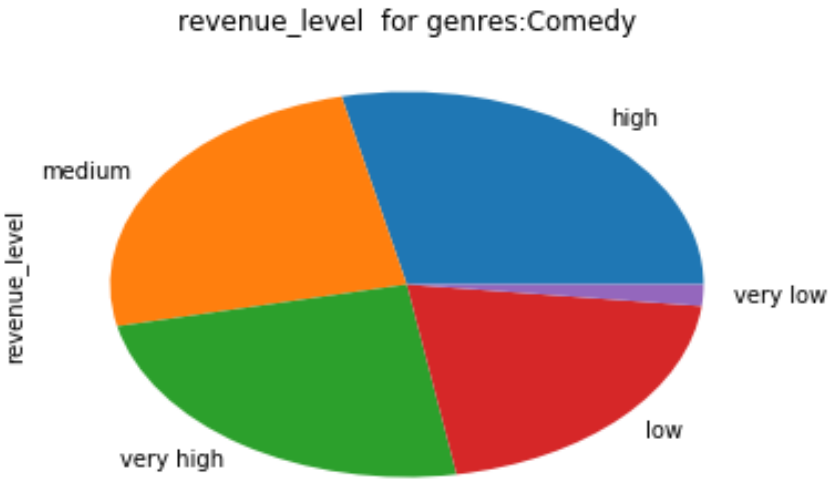
当前所有有效电影风格统计如下：

电影风格	电影数量	备注
Drama	1358	
Comedy	1063	
Thriller	1028	
Action	925	
Adventure	677	
Crime	557	
Romance	496	
Science Fiction	455	
Horror	376	
Family	367	
Fantasy	355	
Mystery	288	
Animation	186	
History	109	
War	101	
Music	99	
Western	39	
Documentary	15	
Foreign	1	

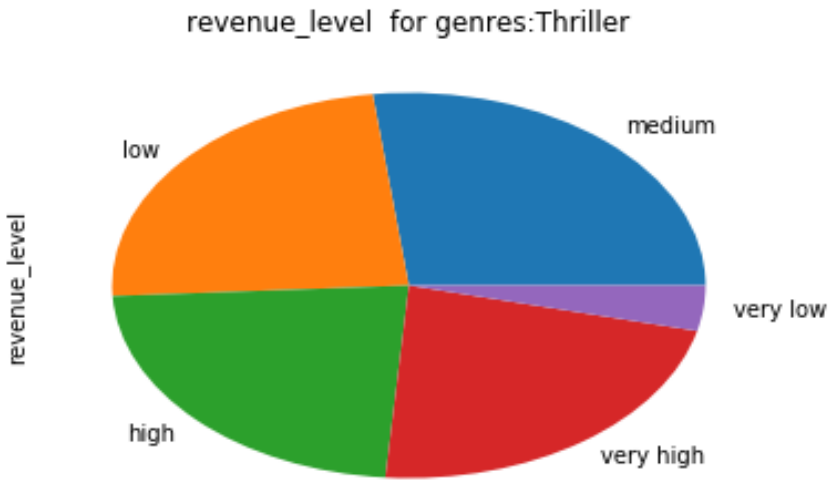
只统计拍摄数量超过 100 部的电影风格，如下：
戏剧类：



喜剧类：



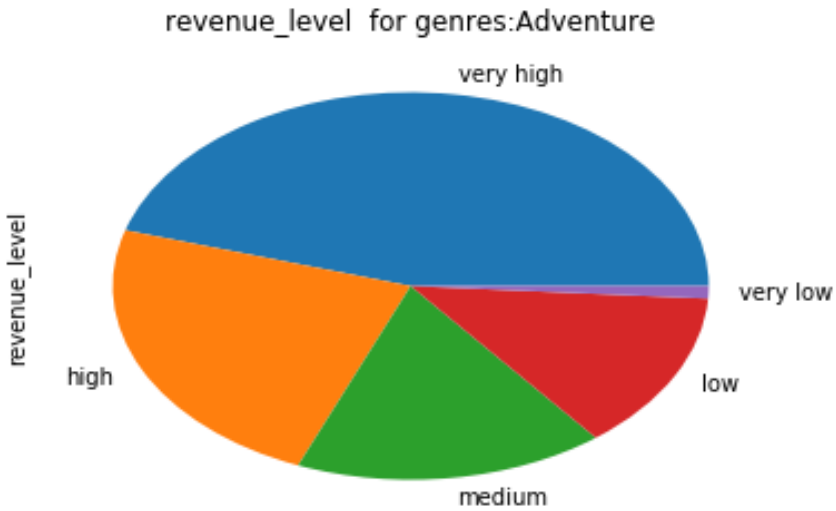
惊悚类：


revenue_level for genres:Thriller
medium
low
very low
revenue_level
high
very high

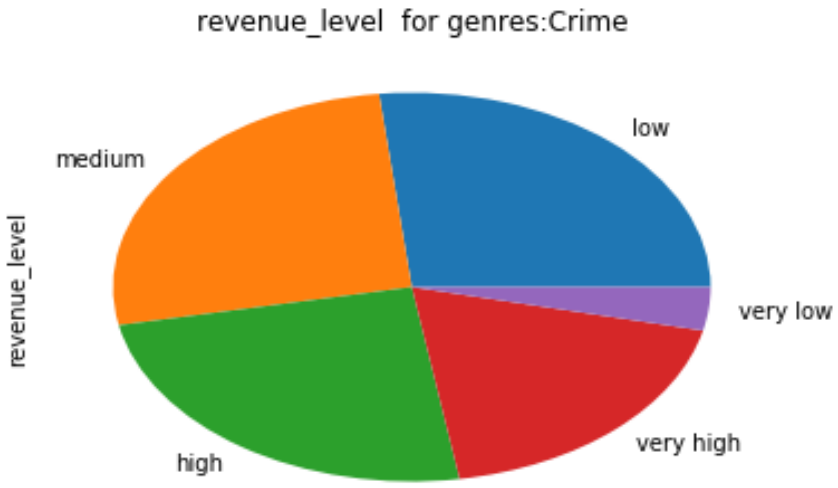
动作类：


revenue_level for genres:Action
very high
high
very low
revenue_level
low
medium

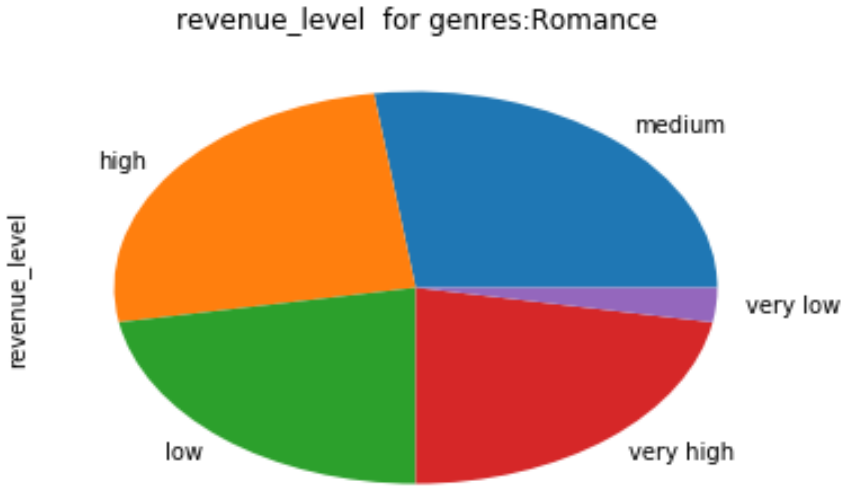
冒险类：


revenue_level for genres:Adventure
very high
very low
revenue_level
high
low
medium

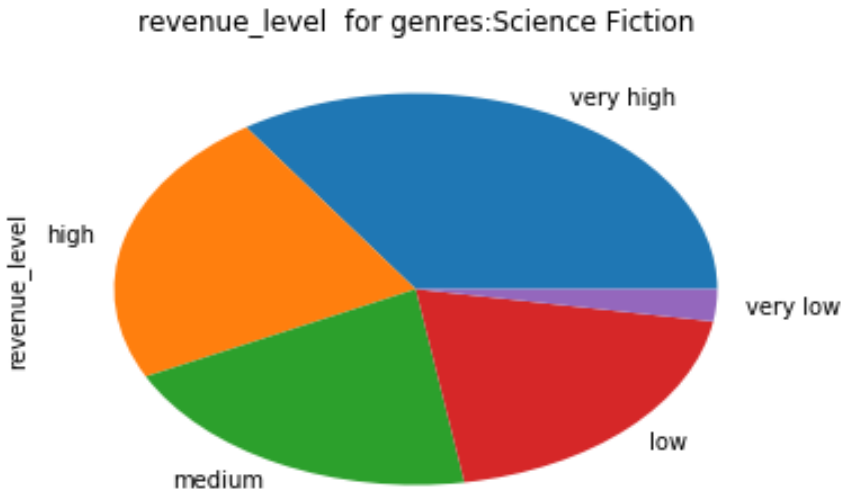
犯罪类：


revenue_level for genres:Crime
low
medium
revenue_level
very low
very high
high

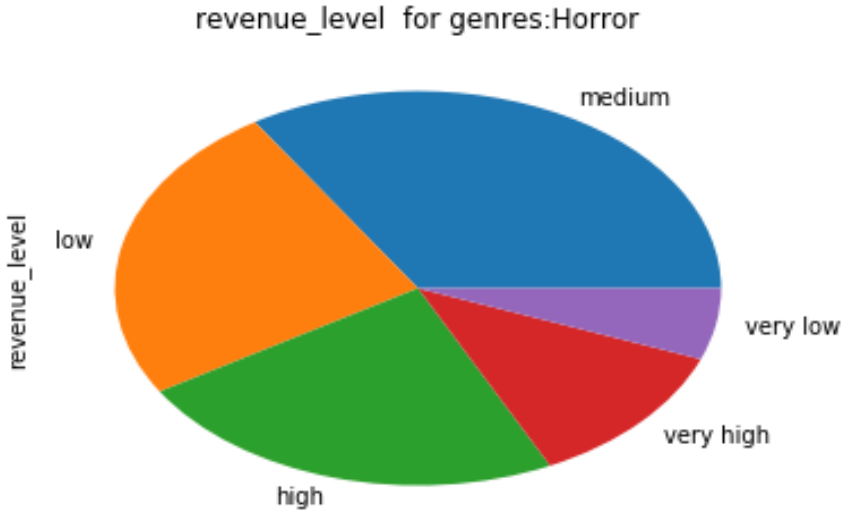
浪漫类：


revenue_level for genres:Romance
medium
high
revenue_level
very low
low
very high

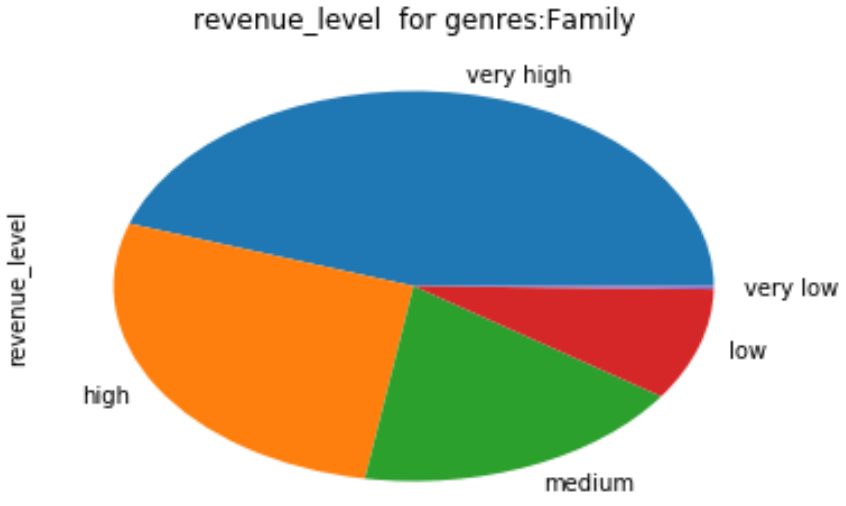
科幻类：


revenue_level for genres:Science Fiction
very high
high
revenue_level
very low
low
medium

恐怖类：


revenue_level for genres:Horror
medium
low
revenue_level
very low
very high
high

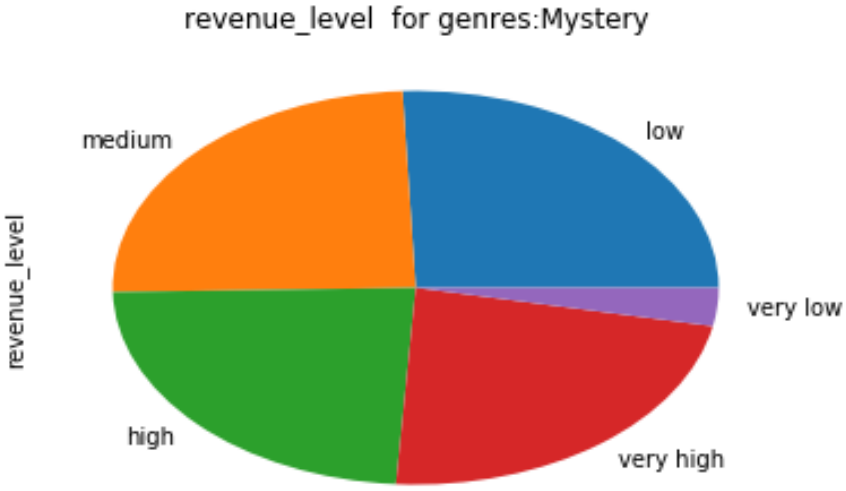
家庭类


revenue_level for genres:Family
very high
revenue_level
very low
low
high
medium

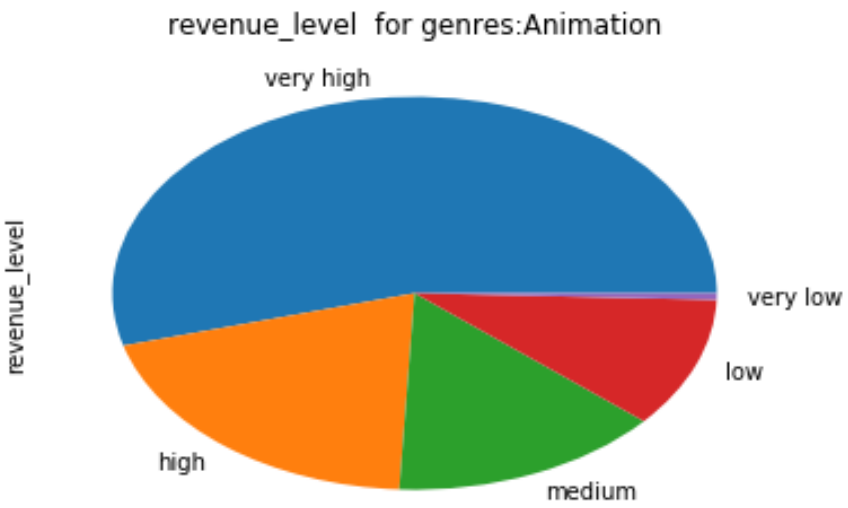
奇幻类


revenue_level for genres:Fantasy
very high
revenue_level
high
very low
low
medium

神秘类


revenue_level for genres:Mystery
medium
low
revenue_level
very low
high
very high

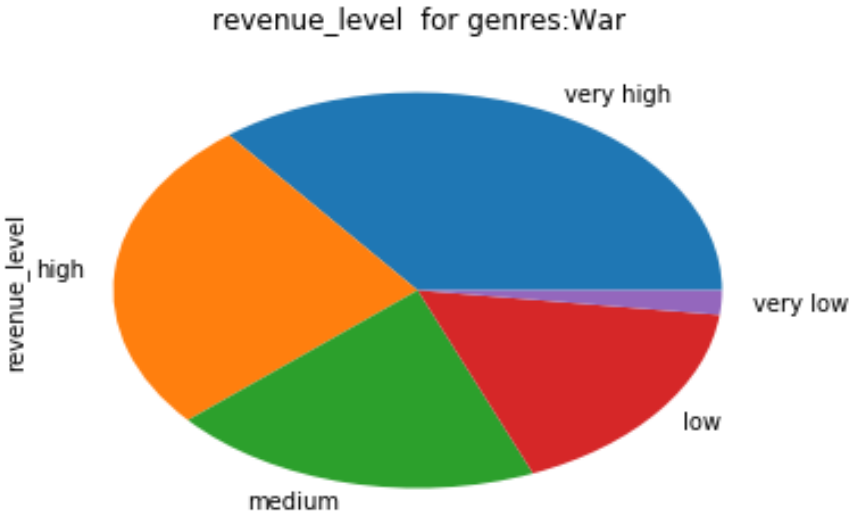
动画类


revenue_level for genres:Animation
very high
very low
low
high
medium
revenue_level

历史类


revenue_level for genres:History
high
very low
low
very high
medium
revenue_level

战争类


revenue_level for genres:War
very high
very low
low
medium
high
revenue_level

2.2.4 结论：

根据本节的分析，得到以下结论：

1. 电影的电影预算对电影平均收入有很大影响；平均受欢迎程度次之；电影的平均评分则对电影收入没有什么贡献；
2. 电影风格的影响：

以下电影风格的收入比较均衡：

戏剧类 喜剧类 惊悚类 犯罪类 浪漫类 神秘类 历史类

以下电影风格有较大概率获得高/极高收入：

动作类 冒险类 科幻类 家庭类 奇幻类 动画类 战争类

以下电影风格有较小概率获得高/极高收入：

恐怖类

2.3 列出净利润最高的前 5 部电影（电影收入-预算最高）

不考虑通货膨胀因素：

排名	电影名称	中文译名	利润（亿美元）	备注
1	Avatar	阿凡达	25.45	
2	Star Wars: The Force Awakens	星球大战：原力觉醒	18.68	
3	Titanic	泰坦尼克号	16.45	
4	Jurassic World	侏罗纪世界	13.64	
5	Furious 7	速度与激情 7	13.16	

考虑通货膨胀因素：

排名	电影名称	中文译名	利润（亿美元）	备注
1	Star Wars	星球大战	27.5	
2	Avatar	阿凡达	25.86	
3	Titanic	泰坦尼克号	22.34	
4	The Exorcist	驱魔人	21.28	
5	Jaws	大白鲨	18.79	

2.4 演员分析；

2.4.1 分析

演员表中包括多个演员，用”|”分隔。

创建一个新数据，只包含“演员”、“平均评分”、“电影收入”、“欢迎程度”，其中每个含 n 个演员的行在新数据中对应 n 行，每行只含 1 个演员。

根据经验，我认为：

一个演员出演电影的数量能间接反映这个演员的演技和口碑，优秀的演员能够得到很多的出演机会；

一部电影的平均评分反映了这部电影的艺术水平，间接反映主演演员的演技；

一部电影的受欢迎程度反映了这部电影的人气，间接反映主演演员的人气；

一部电影的电影收入则取决于很多因素：题材、欢迎程度、投入、宣传等，演员在其中的影响并不是非常重要（一个经常出演科幻和动作大片的演员，其作品的平均收入毫无疑问地碾压经常出演文艺电影的演员，但这不能说明他的水平更高）。

综上，我们从出演电影的数量、平均评分和平均受欢迎程度这 3 个角度来分析一个演员的演技、人气和口碑。

2.4.2 演员参演电影数量分布

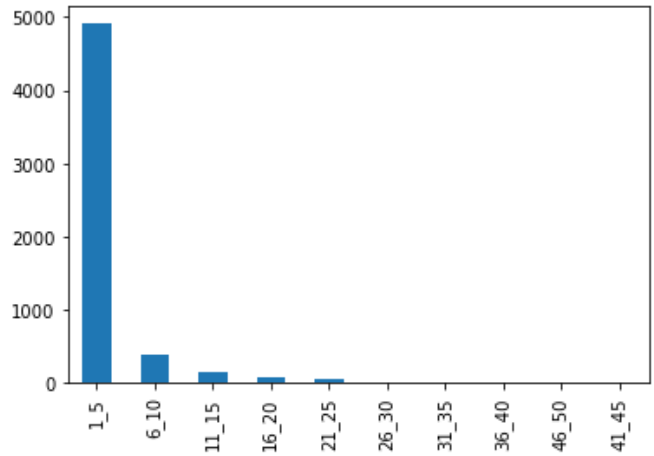
将上述数据中“演员”一列单独取出，通过 value_count 函数可以统计每个演员出演的电影数量。分析结果发现，出演最多的演员出演了 50 部（只有 Robert De Niro 1 人），最少只有 1 部（很多人）。

我们将数量按 5 为间隔分成 1-5 6-10 …… 41-50 10 个等级，统计每一等级的出演人员数量。如下：

演员出演数量	人数
1-5	4908
6-10	383
11-15	146
16-20	68
21-25	43
26-30	13
31-35	7
36-40	2
41-45	1

46-50	1

一个形象的示意图如下：



可以看出，绝大多数演员参演的电影数量少于 5 部，只有少数演员参与了很多电影的演出。

考虑到任何一部电影都由很多演员出演，单一演员的影响有很强的随机性…而我们的目的是分析演员本身的演艺水平。因此我们在分析过程中排除在数据中出演电影数量不超过 5 部的演员，以避免统计偏差。

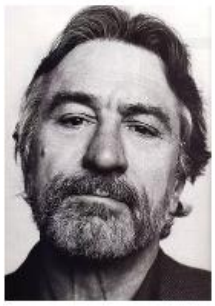
2.4.3 参演电影数量最高的前 5 名演员

我们首先统计在数据库中出演电影数量最多的演员，如下：

排名	演员	出演电影数量
1	Robert De Niro	50
2	Bruce Willis	43
3	Samuel L. Jackson	40
4	Nicolas Cage	40
5	Johnny Depp	35

分别是以下几位大神：

罗伯特·德尼罗 Robert De Niro



性别: 男
星座: 狮子座
出生日期: 1943-08-17
出生地: 美国,纽约
职业: 演员 / 制片 / 配音 / 导演
更多外文名: Robert Mario De Niro Jr
Robert Mitchum / Bob(昵称)
更多中文名: 罗拔·迪尼路 / 劳勃·狄尼
家庭成员: Grace Hightower（妻） /

布鲁斯·威利斯 Bruce Willis

性别: 男
星座: 双鱼座
出生日期: 1955-03-19
出生地: 德国,伊达尔-奥伯斯坦
职业: 演员 / 制片 / 编剧
更多外文名: Walter Bruce Willis(本名) / Br
Willis(昵称)
更多中文名: 布斯·韦利士 / 布鲁斯·威利
家庭成员: 黛米·摩尔(前妻) / 艾玛·赫明(妻)

塞缪尔·杰克逊 Samuel L. Jackson

性别: 男
星座: 射手座
出生日期: 1948-12-21
出生地: 美国,华盛顿,哥伦比亚特区
职业: 演员 / 配音 / 制片
更多外文名: Samuel Leroy Jackson (本名) /
(昵称) / Sam (昵称)
更多中文名: 森姆·积逊(港)
家庭成员: LaTanya Richardson(妻)

尼古拉斯·凯奇 Nicolas Cage

性别: 男
星座: 摩羯座
出生日期: 1964-01-07
出生地: 美国,加利福尼亚,长滩
职业: 演员 / 制片 / 配音 / 导演
更多外文名: Nicholas Kim Copp
更多中文名: 尼古拉斯·基治 / 尼
家庭成员: 弗朗西斯·福特·科波拉
imdb编号: nm0000115

约翰尼·德普 Johnny Depp

增改描述、换头像

性别: 男
星座: 双子座
出生日期: 1963-06-09
出生地: 美国,肯塔基,欧文斯
职业: 演员 / 制片 / 配音 / 导
更多外文名: John Christoph
(昵称)
更多中文名: 尊尼狄普 / 强尼
家庭成员: John Christopher
Heard(前妻) / Christie Demb

分析：可以看出，参演电影数量是一个衡量演员人气、演技和口碑的综合指标的不错方式：罗伯特德尼罗，演技之神，毫无争议的影帝，一生塑造的经典形象数不胜数；布鲁斯威利斯，动作片大神，因《虎胆龙威》闻名于世，电影史上最知名的硬汉；塞缪尔杰克逊，好莱坞最具实力的黑人男星；尼古拉斯凯奇，30年的演艺经历，《石破天惊》《战争之王》《勇闯夺命岛》等一系列高水准高口碑的电影，早已奠定了他的地位；强尼德普，永远的杰克船长，《加勒比海盗》系列的绝对主角，此外还有《理发师陶德》《爱丽丝梦游仙境》等一批经典电影的演艺经历。

其他在数据中出演了超过 30 部的电影的演员的名字：
布拉德•皮特：经典电影 《十一罗汉》系列 《特洛伊》《world war z》
马特•达蒙: 经典电影《拯救大兵瑞恩》《星际穿越》《火星救援》
西尔维斯特 • 史泰龙：经典电影《第一滴血》系列 《敢死队》系列
汤姆•克鲁斯：经典电影《碟中谍》系列
摩根•弗里曼：好莱坞的黄金配角，经典电影《肖申克的救赎》、《七宗罪》、《百万美元宝贝》、《蝙蝠侠前传》系列
汤姆汉克斯：经典电影《阿甘正传》、《拯救大兵瑞恩》《达芬奇密码》

2.4.4 参演电影平均评分最高的 5 位演员

统计所有演员的平均评分后，结果如下：

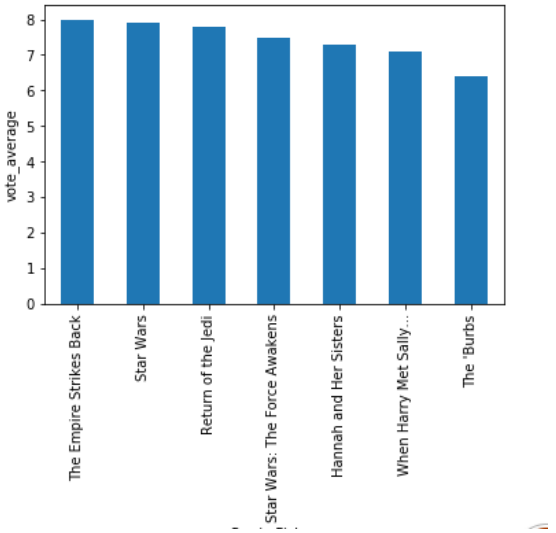
排名	演员	评分值
1	Carrie Fisher	7.43
2	Rupert Grint	7.2
3	Leonardo DiCaprio	7.06
4	Faye Dunaway	7.036
5	Alan Rickman	7.01

检索到了 5 名演员的资料，以及他们拍摄的所有电影的评分列表图

凯丽·费雪 Carrie Fisher

性别: 女
星座: 天秤座
生卒日期: 1956-10-21 至
出生地: 美国,加利福尼亚
职业: 演员 / 配音 / 编剧
更多外文名: Carrie Franc
更多中文名: 凯莉·费雪
家庭成员: Debbie Reyno
imdb编号: nm0000402
官方网站: http://www.ca

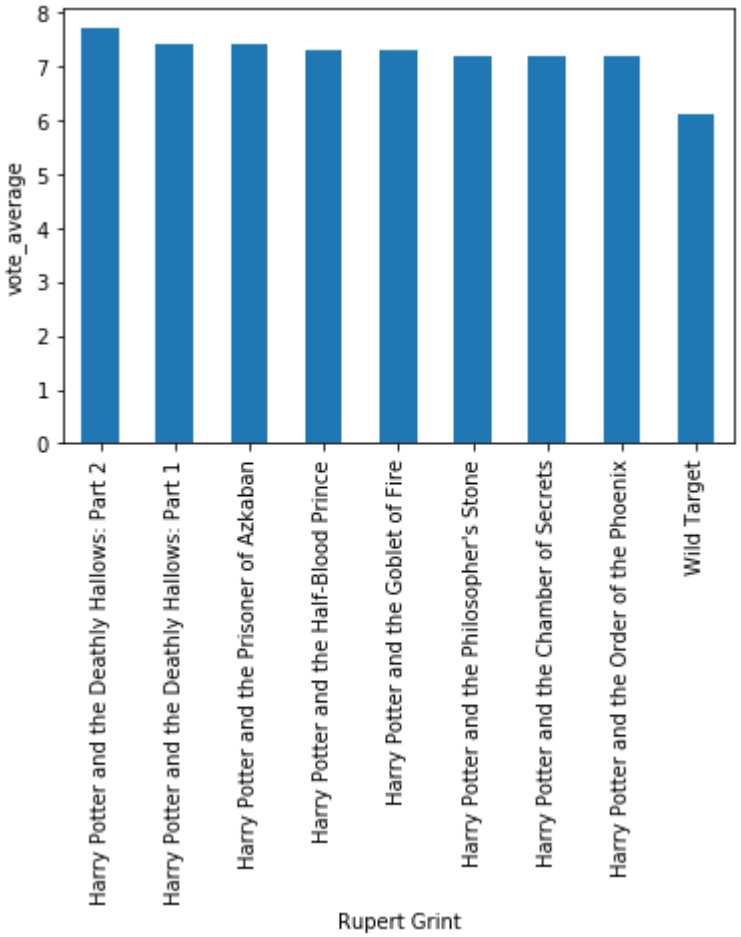
增改描述、换头像



鲁伯特·格林特 Rupert Grint

性别: 男
星座: 处女座
出生日期: 1988-08-24
出生地: 英国,赫特福德郡,斯蒂夫
职业: 演员
更多外文名: Rupert Alexander L
更多中文名: 鲁伯特·亚历山大·
imdb编号: nm0342488
官方网站: http://www.rupert-grir

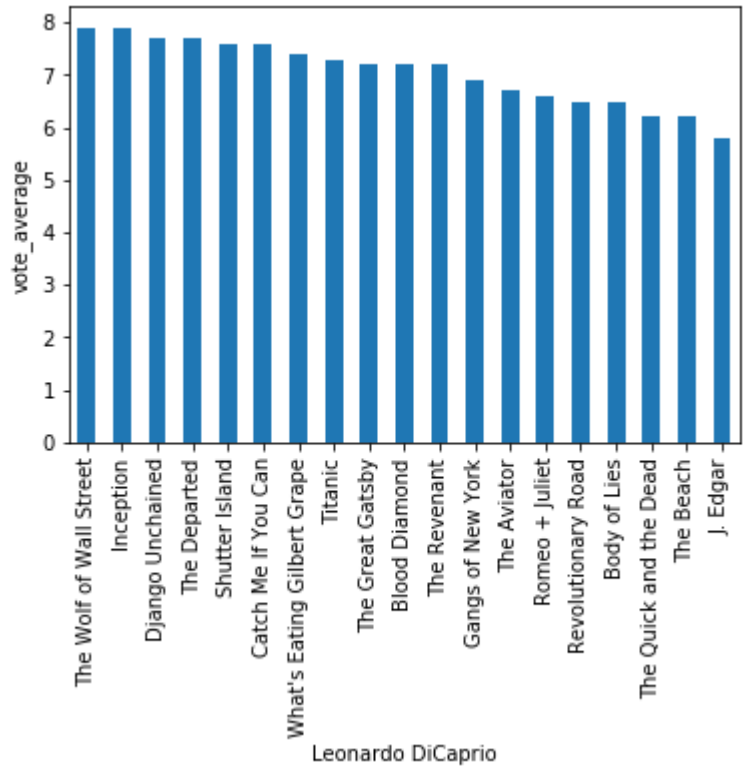
增改描述、换头像



莱昂纳多·迪卡普里奥 Leonardo DiCaprio

增改描述、换头像

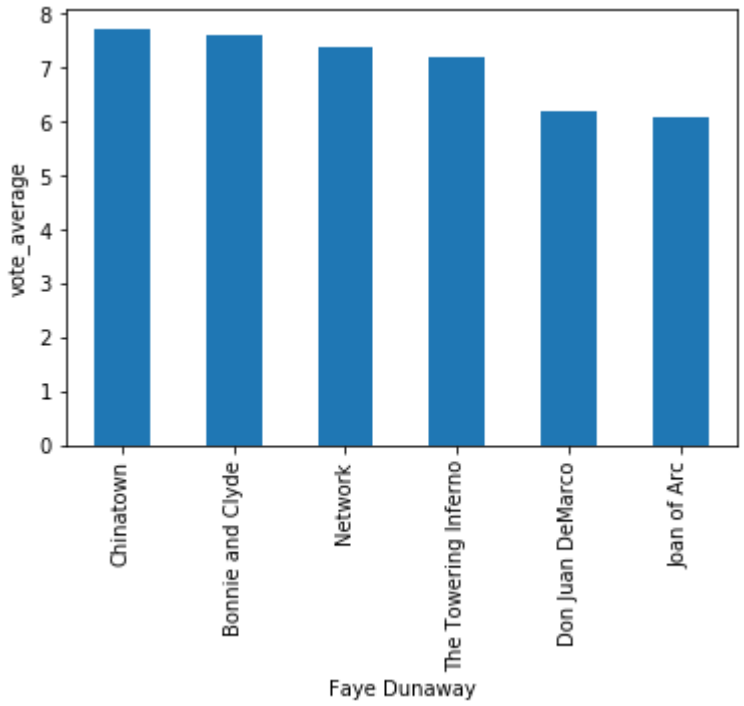
性别: 男
星座: 天蝎座
出生日期: 1974-11-11
出生地: 美国,加利福尼亚,洛杉矶
职业: 演员 / 制片 / 编剧 / 配音
更多外文名: Leonardo Wilhelm DiCaprio (本名) / Lenny D (
更多中文名: 李奥纳多·迪卡普里奥 / 里安纳度·迪卡比奥 / 小
(昵称)
家庭成员: Gisele Bündchen(前女友) / Bar Refaeli(前女友) /
Heatherton(前女友) / Kelly Rohrbach(前女友)



费·唐纳薇 Faye Dunaway

性别: 女
星座: 摩羯座
出生日期: 1941-01-14
出生地: 美国,佛罗里达州,巴斯科姆
职业: 演员 / 制片 / 编剧 / 导演
更多外文名: Dorothy Faye Dunaway (本
家庭成员: Terry O'Neill（前夫） / Peter
imdb编号: nm0001159

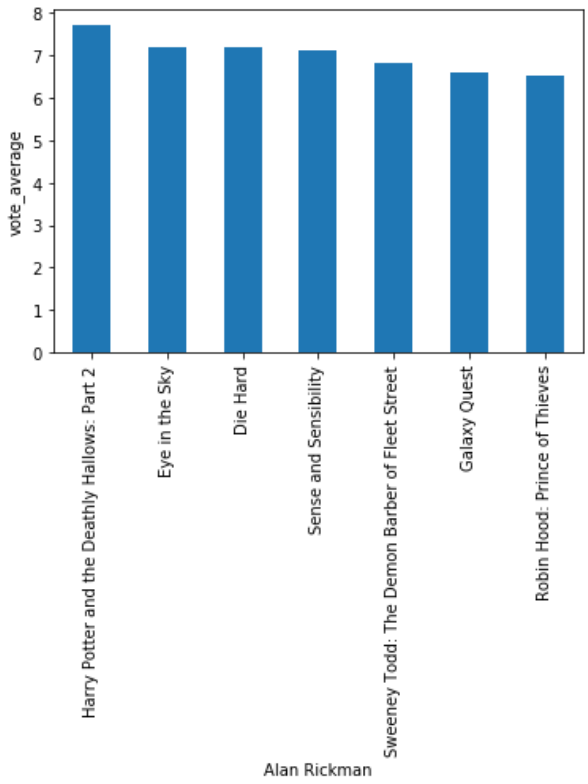
增改描述、换头像



艾伦·瑞克曼 Alan Rickman

增改描述、换头像

性别: 男
星座: 双鱼座
生卒日期: 1946-02-21 至 2016-
出生地: 英国,伦敦,哈默史密斯
职业: 演员 / 导演 / 配音 / 编剧
更多外文名: Alan Sidney Patric
更多中文名: 艾伦·里克曼 / 阿伦
家庭成员: Rima Horton(妻)
imdb编号: nm0000614



分析：

从上面的分析中看出，出演电影平均评分最高的演员可以分为下述两类：

1、出演了一个评分很高的、特定的电影系列

典型为 Carrie Fisher、Rupert Grint。二者的平均评分高主要是源于各自出演的《星球大战》系列（出演莱雅公主）和《哈利•波特》（出演罗恩）系列的高评分。他们在这个数据库中其它电影的评分就出现了明显的反差。Carrie Fisher 的星战 4 部电影最低的平均评分也是 7.5，而其它 3 部电影在 6.4-7.3 之间（仍然是不低的评分）。而 Rupert Grint 更明显，其出演的 9 部电影，其中 8 部是《哈利•波特》系列，均在 7.2 分以上，唯一一部非该系列的电影就只有 6.1 了。显然，他们的上榜更多是沾了电影系列的光彩，不能完全反映真实演技。

此外，顺便调查了这两个系列的另外两个主角：卢克•天行者的扮演者 Mark Hamill 和哈利•波特的扮演者 Daniel Radcliffe 的情况。Mark Hamill 的平均评分甚至比这里的 5 名演员更高，但因为在数据库中只有 5 部电影（星战 4 部+蝙蝠侠 1 部）被排除了。Daniel Radcliffe 则是因为出演的其它 3 部电影评分偏低（分别是 6.7 6.1 5.5），导致平均分被拉低。

2、演员本身出演的电影大部分评分较高

这一类演员可以说主要是凭借自身的高水平保证了其出演的大部分电影都有着较高的口碑。典型代表是排第 3 名的莱昂纳多，莱昂纳多出演了 19 部不同的电影，其中接近 2/3 是 6.9 分以上的高分电影，只有 1 部未超过 6 分。评分最高的两部电影（《华尔街之狼》和《盗梦空间》）都达到 7.9。本人最熟悉的《泰坦尼克号》和《禁闭岛》在他的评分榜单上只能排到第 4 和第 7。这 5 名演员中除他之外其它人在这个数据库中最多也只出演了 9 部电影（Rupert Grint，其中 8

部是《哈利波特》系列，见上），不到他的一半。可以这么说，从出演电影的平均评分统计结果来看，**莱昂纳多是当之无愧的演技之王，是电影艺术质量的绝对保证**。

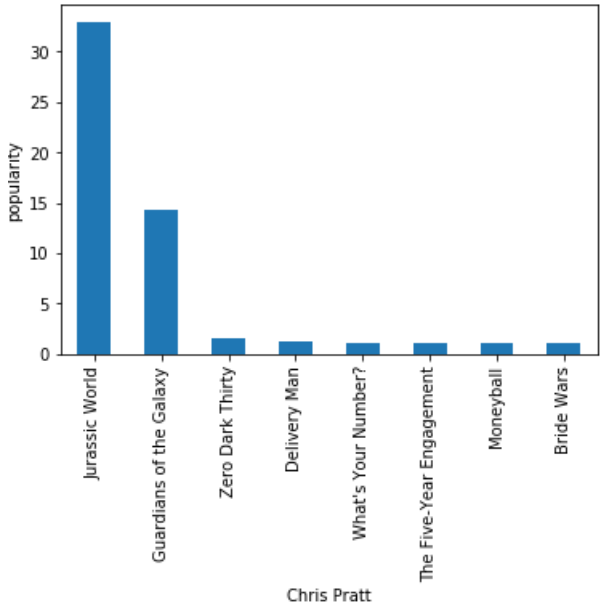
排在第 4 和第 5 的 Faye Dunaway 和 Alan Rickman 在数据库中分别出演了 6 部和 7 部电影。Faye Dunaway 是老牌影星，但她的巅峰时期在 60-70 年代，多数数据不全。在这里排名较高似乎可能是收录在最终有效数据中的恰好是她评分较高的那些电影，有幸存者偏差的嫌疑。Alan Rickman 可能有类似的成分，收录在本数据库中的都是他出演的非常知名的电影：《哈利波特》《天空之眼》《虎胆龙威》《理智与情感》，都是评分在 7 分上的高口碑电影，同时其它 3 部也都在 6.5 分以上，没有拉低平均值。这一排名能够证明他们 2 人的演技水平很高，但不能完全证明他们的地位是否可以达到前 5 水准。

2.4.5 参演电影平均受欢迎程度最高的 5 位演员

统计所有演员的平均受欢迎程度，如下：

排名	演员	平均受欢迎程度（%）
1	Chris Pratt	6.82
2	Bryce Dallas Howard	6.78
3	Nicholas Hoult	5.86
4	Tom Hardy	5.76
5	Carrie Fisher	5.37

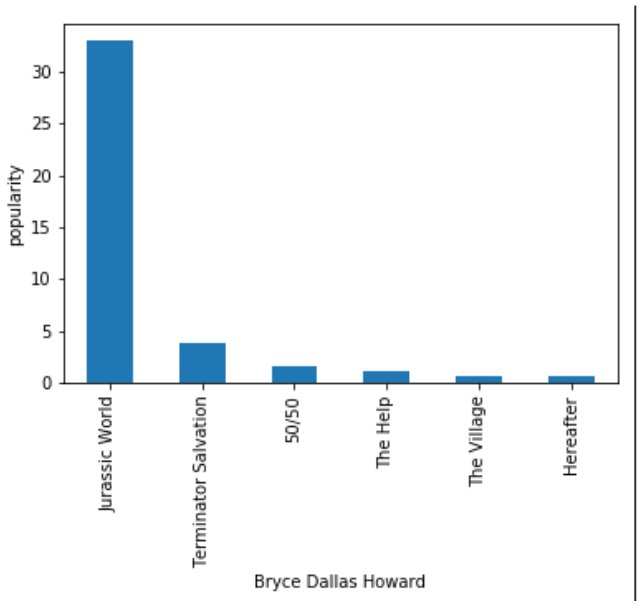
检索到了 5 名演员的资料，以及他们拍摄的所有电影的受欢迎程度列表图



布莱丝·达拉斯·霍华德 Bryce Dallas Howard

增改描述、换头像

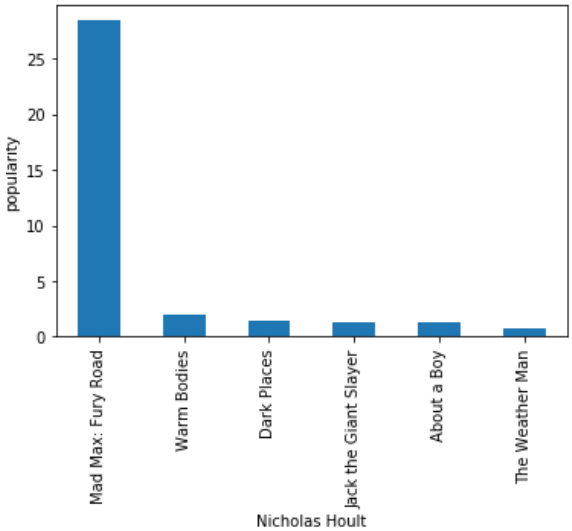
性别: 女
星座: 双鱼座
出生日期: 1981-03-02
出生地: 美国,加利福尼亚州,洛杉矶
职业: 演员 / 导演 / 编剧 / 配音
更多外文名: Bry (昵称)
更多中文名: 比丝多丽丝候活
家庭成员: 朗·霍华德(父) / 塞斯·盖贝尔(夫)
imdb编号: nm0397171



尼古拉斯·霍尔特 Nicholas Hoult

性别: 男
星座: 射手座
出生日期: 1989-12-07
出生地: 英国,伯克郡,沃金厄姆
职业: 演员 / 配音
更多外文名: Nicholas Caradoc Hoult (本
更多中文名: 尼古拉斯·侯特(港) / 尼克拉斯
imdb编号: nm0396558
官方网站: www.nicholashoult.net

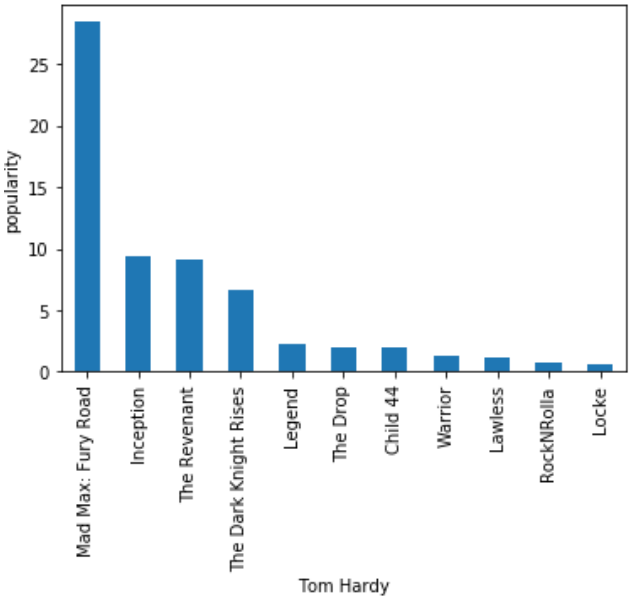
增改描述、换头像



汤姆·哈迪 Tom Hardy

性别: 男
星座: 处女座
出生日期: 1977-09-
出生地: 英国,伦敦,
职业: 演员 / 制片 /
更多外文名: Edwar
更多中文名: 汤老湿
家庭成员: Chips Ha
imdb编号: nm0362

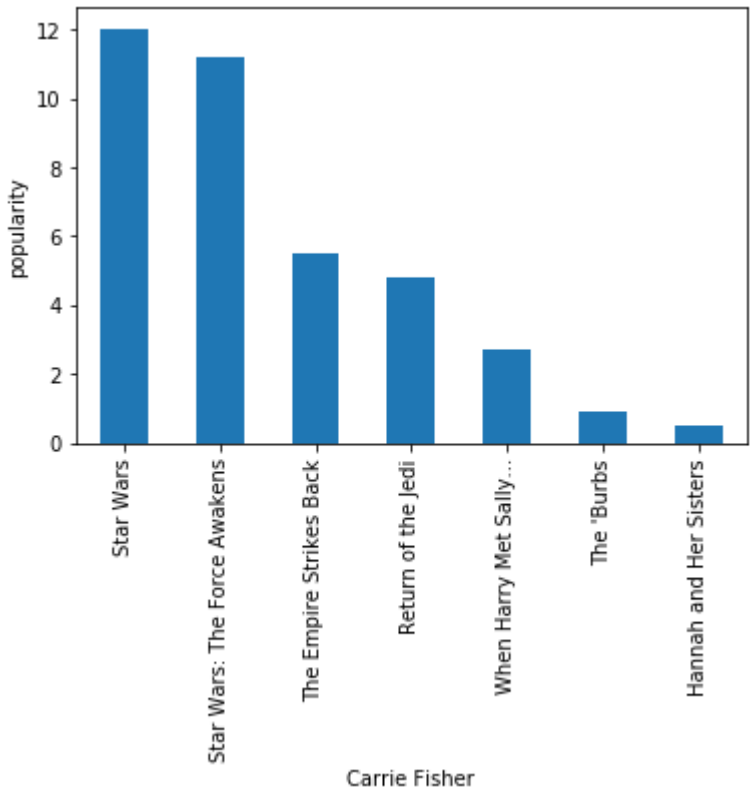
增改描述、换头像



凯丽·费雪 Carrie Fisher

性别: 女
星座: 天秤座
生卒日期: 1956-10-21 至
出生地: 美国,加利福尼亚
职业: 演员 / 配音 / 编剧
更多外文名: Carrie Franc
更多中文名: 凯莉·费雪
家庭成员: Debbie Reyno
imdb编号: nm0000402
官方网站: http://www.ca

增改描述、换头像



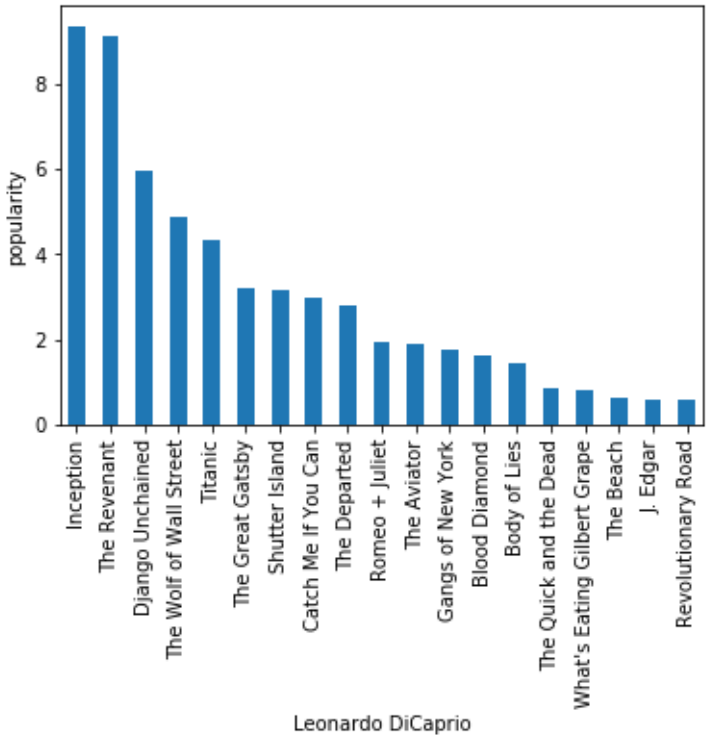
从上述数据中可以看出：

每个演员出演的电影的受欢迎程度差别极大，说明预判存在问题，电影的受欢迎程度与每个演员的人气没有较大的关系。

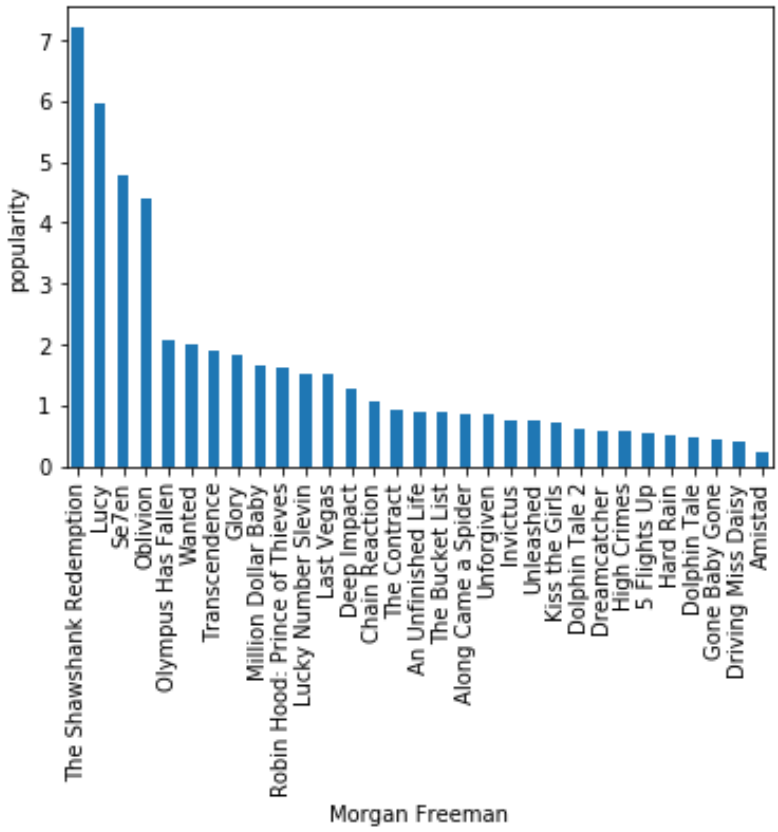
进一步仔细观察就可以发现，上述 5 名演员中，凯丽·费雪的平均受欢迎程度毫无疑问完全来自 4 部星球大战的极高人气。Bryce Dallas Howard 和 Nicholas Hoult 各自只靠一部电影的超高人气支撑（《侏罗纪世界》和《疯狂麦克斯》），Chris Pratt 主要依靠《侏罗纪世界》和《银河护卫队》。Tom Hardy 表现稍微优异一些，虽然《疯狂麦克斯》仍然对他的排名起了关键的作用，但至少他还有《盗梦空间》、《荒野猎人》和《蝙蝠侠：黑暗骑士崛起》这 3 部人气不错的电影。但是考虑到他在这 3 部电影中都不是人气主要贡献者（值得注意的是，这 3 部有 2 部主演是莱昂纳德），因此他的人气地位也很值得怀疑。

作为对比，我们选取了 2 位毫无争议（我最喜欢）的优秀演员：莱昂纳多和摩根弗里曼的数据进行对比：

莱昂纳多



摩根弗里曼



可以发现，莱昂纳多和摩根弗里曼出演的电影在受欢迎程度方面也有非常大的反差。从不到 1%到 8%都有。这使得他们的平均受欢迎程度无法与 Chris Pratt、

Bryce Dallas Howard、Nicholas Hoult、Tom Hardy 这些人相比，这些人每人都出演了一部受欢迎程度达到 30%以上的电影，其它电影则要么数量不多，要么受欢迎程度也不足够低，无法将平均程度拉低足够多。

结论：演员参演电影的平均受欢迎程度与该演员的人气没有明显的关系，排名靠前的都是那些恰好参与了一部或几部高人气电影，同时其他作品又很少的演员。造成这一点有 2 个原因：一是数据记录不全，每个演员参演的所有电影没有都列在数据库中；二是数据中只显示某个电影中有哪些演员参演，而不会区分这些演员在电影中的重要性。

2.4.6 结论

本轮对数据库中出现的所有演员进行了分析，分别从出演电影数量、出演电影平均评分、出演电影平均受欢迎程度这 3 个角度考察了演员的演技。结论如下：

出演电影数量能够基本代表演员在业界的地位和综合实力；出演电影平均评分能够在一定程度上代表演员的演技，但是那些出演了一些非常受欢迎的电影系列的演员在此项统计中会得到优势；出演电影的平均受欢迎程度对演员的人气没有太大的说服力。

2.5 观察电影平均预算和收入是否随着发行年代上涨；

2.5.1 分析内容

将电影按发行年代以 10 年为周期，计算每个周期内电影的平均预算和平均收入。为了能够更好地计算统计规律，排除个别电影对统计值的干扰，将周期和周期之间设置 5 年的重叠期。

2.5.2 分析方法

分析数据中的电影发行年代，最早为 1960 年，最晚为 2015 年。且早期（60 年代-70 年代初）每年的电影基本都是个位数，不具备统计价值。故将分析的时间起点设为 1976 年，分段如下：

电影年代名称	发行时间	备注
Old	1976 前	不统计

1980s	1976-1985	
1985s	1981-1990	
1990s	1986-1995	
1995s	1991-2000	
2000s	1996-2005	
2005s	2001-2010	
2010s	2006-2015	

在数据中新增“电影年代”，年代的值采用上表中的“命名”列。通过该列的值将数据分组，计算每组的平均预算和平均电影收入。

由于采用了滑窗划分，故多数电影都会处于 2 个年代，难以直接应用 groupby 方法（这种分组是不重叠的）。解决方法是采用奇偶划分法：

奇分组（称为“电影年代 1”）

电影年代名称	发行时间
Old	1976 前
1980s	1976-1985
1990s	1986-1995
2000s	1996-2005
2010s	2006-2015

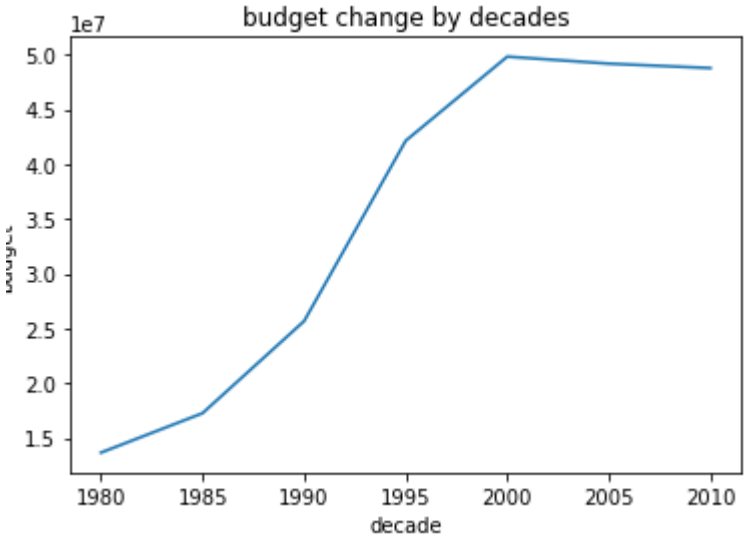
偶分组（称为“电影年代 2”）

电影年代名称	发行时间
Old	1981 前
1985s	1981-1990
1995s	1991-2000
2005s	2001-2010

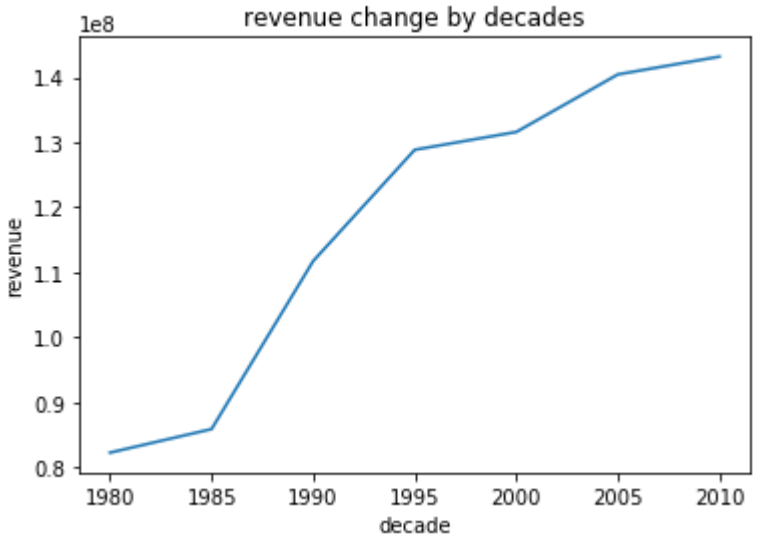
分别计算各组均值，然后将两组均值组合在一起。

2.5.3 分析结果

各个年代电影平均预算统计结果：



各个年代电影平均收入统计结果：



2.5.4 分析结论

证明电影平均预算和平均收入均在随着时间上涨。注意到预算上涨了三倍多，而平均收入只增长了不到一倍。

3 总结

3.1 完成的工作

本报告根据所提供的 TMDb 数据，进行了以下工作：

1. 对数据的有效性进行了检查，并根据评价次数、预算和电影收入，剔除了一部分可疑数据；
2. 从以下三个角度对数据进行了分析：
   1） 高收入电影的主要特点；
   2） 哪些特征可以用于评价演员的演艺水平；
   3） 不同年代的电影平均预算和收入变化情况；

3.2 分析存在的局限性

3.2.1 数据样本量不大，分布不平衡

原始数据有 1 万多部电影的信息，但在剔除无效和可疑数据后，还剩余 3126 部电影，这使得数据样本的数量不能达到大数据的量级；在分析演员水平时，这一问题导致了很大的影响。

此外，很多特征的数据分布极不均衡，例如，不同电影风格的电影数量分布就极不均衡，导致一些风格的电影由于样本数量太少，无法分析其对应的收入特点。

3.2.2 电影预算和收入的正确性存在疑问

在数据中有大量极不正常的预算/收入值（主要是偏低）。例如，原数据中有多达 5000 多条记录的预算值是 0！

即使排除了预算值和收入小于 100 美元的记录，剩下的数据仍然存在明显不正常的情况。例如，1986 年发行的电影《The Karate Kid, Part II》（译《龙威小子 2》），预算数据为 113 美元，票房却达到 1.15 亿美元（赚翻了，微电影也能有大市场？？），完全违背常理。显然预算数据是错误的。

另一例是 2003 年发行的电影《brother bear》，预算达 1 亿美元，而收入为 250 美元（简直是史上亏损率最高的电影）。实际查询得知，该片票房达 2.26 亿美元。

可以看出该数据中的电影预算和收入统计值的可信度不高，存在为数不少的错误记录。由于像上面这样明显的错误不可能很多，因此将错误数据全部挑出是不可能的。这不可避免地会影响相关分析的准确度，如对高收入电影特点的分析。

3.2.3 演员信息不能完全反映演员的演艺水平

分析中已经提到，本数据中与演员相关的部分，存在两个严重问题：

1. 收录不全

多数演员都只有一部分电影收录在这份数据中（特别是那些一生中多数电影都不很知名的演员，收录在这里的电影只占其出演电影的一小部分）；

2. 未体现不同演员在同一电影中的重要性

每个电影中多个演员完全并列，不能体现演员对电影的重要性差异

由于上述原因的存在，通过电影情况对演员情况进行分析的准确度受到了很大影响，如前所述，在利用出演电影的平均特征来评价演员时，那些出演了很多经典作品的影帝级人物，往往不如一个恰好出现在几部成功电影中的普通演员—即使后者在那些电影中是并不重要的配角。这些普通演员出演的其它并不成功的电影根本没有出现在所分析的数据中。

4 引用

本报告所引用数据以外的电影和演员信息均来自豆瓣的公开资料。