**基于书籍评分和阅读量的高分作者特质分析**

**摘要：**如今不少人希望在写作领域取得成就，一位优秀作者的特质是值得每个人学习的。本课题基于“一万本好书的评分数据数据集”，通过数据可视化的手段，从读者的从众心理、书籍评分与阅读量、作者入选书籍数量与作品评分稳定性等方面展开研究，总结优秀作者的共同点。其中，本课题通过对网络上搜集到的29位作者的样本的研究，找到了计量作者所获成就大小的最佳标准；通过各作者书籍的阅读量平均值、评分平均值的分布情况，得出读者倾向于挑选优秀作家的书籍进行阅读，但对书籍的评分较为客观的结论；通过论证多数作者都同时拥有高、低分作品，以及阅读量比书籍评分对作者更重要，解释了与评分有关的标准为何不宜用来衡量作者成就，并得出结论：好的作者要在大量的写作和试错中成长，不断提升自身水平并努力写出对广大群众有吸引力、有意义的作品。最后，本课题探究了作者的成就与其入选书籍数量、作品评分稳定性的关系，作为佐证和结论深化。

**关键字：**大数据，文学，书籍评分规律，优秀作者特质

# 研究背景

在当下网络文学盛行的时代，很多人都有想要在小说领域大展身手的愿望，但却不是很清楚如何成为一个优秀的作者，并获得一定成就。因此，本课题选择了“一万本好书的评分数据数据集”，对其进行数据分析和研究，希望能够了解到读者挑选书籍，及如何对书籍进行评分的一些规律，以及作者想要获得较大成就所需具备的一些条件。

# 研究问题

通过对“一万本好书的评分数据数据集”的统计分析得出，读者挑选书籍，和对书籍进行评分有何规律？作者要想获得较大的成就需要具备哪些条件？

# 研究设计

首先，寻找计量作者所获成就大小的最佳标准。

接着，本课题进行了一些与研究有关的定义：

1. 将上述计量作者所获成就大小的最佳标准定义为“作者的评分”。
2. 将平均评分≥ 4的书籍称为“高分书籍”，反之称为“低分书籍”。

再者，为研究“读者对书籍评分的方式的一些规律”，本课题进行了以下问题的探索：

读者在挑选书籍和对书籍进行评分两方面，是否受到了对作者的从众心理的影响，即是否受到作者的名气的影响？

为此，本课题设计以下研究内容：

1. 作者各书籍阅读量平均值、评分平均值的分布及偏态情况；
2. 同评分档次，不同作者的低分作品阅读量的对比。

然后，为研究“作者要想获得较大的成就需要具备的条件”，本课题将研究分为两条思路：

1. 从“为何其它作者评分的标准不合理”以及“阅读量和评分哪个对作者更重要”入手，以一条逻辑链研究评分和阅读量的相关内容；

为此，本课题设计以下研究内容：

1. 为了解“各作者书籍评分平均值不宜作为其成就的衡量标准”的原因，统计是否大多数作者都有一些高分作品，也有不少低分作品；
2. 探究作者的成就与其书籍最大评分的关系；
3. 探究作者的成就与其书籍平均阅读量的关系；
4. 对比书籍最大评分和平均阅读量对作者评分的影响程度。
5. 朝各作者选入书籍的数量和发挥的稳定性方向进行研究。

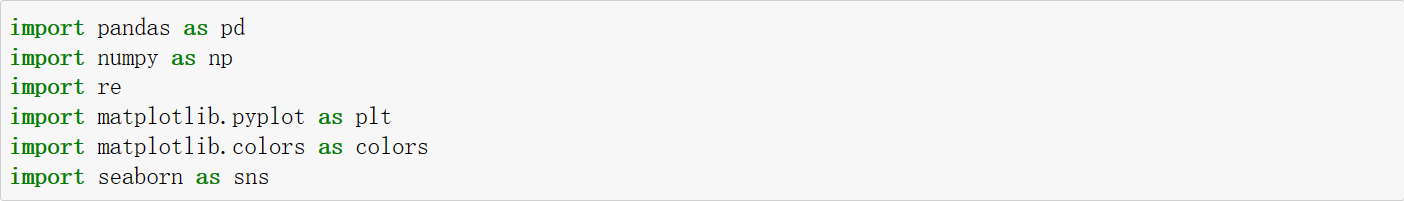
即探究作者的成就与其选入书籍数量、水平发挥的稳定性的关系。

本课题设计了一下研究内容：

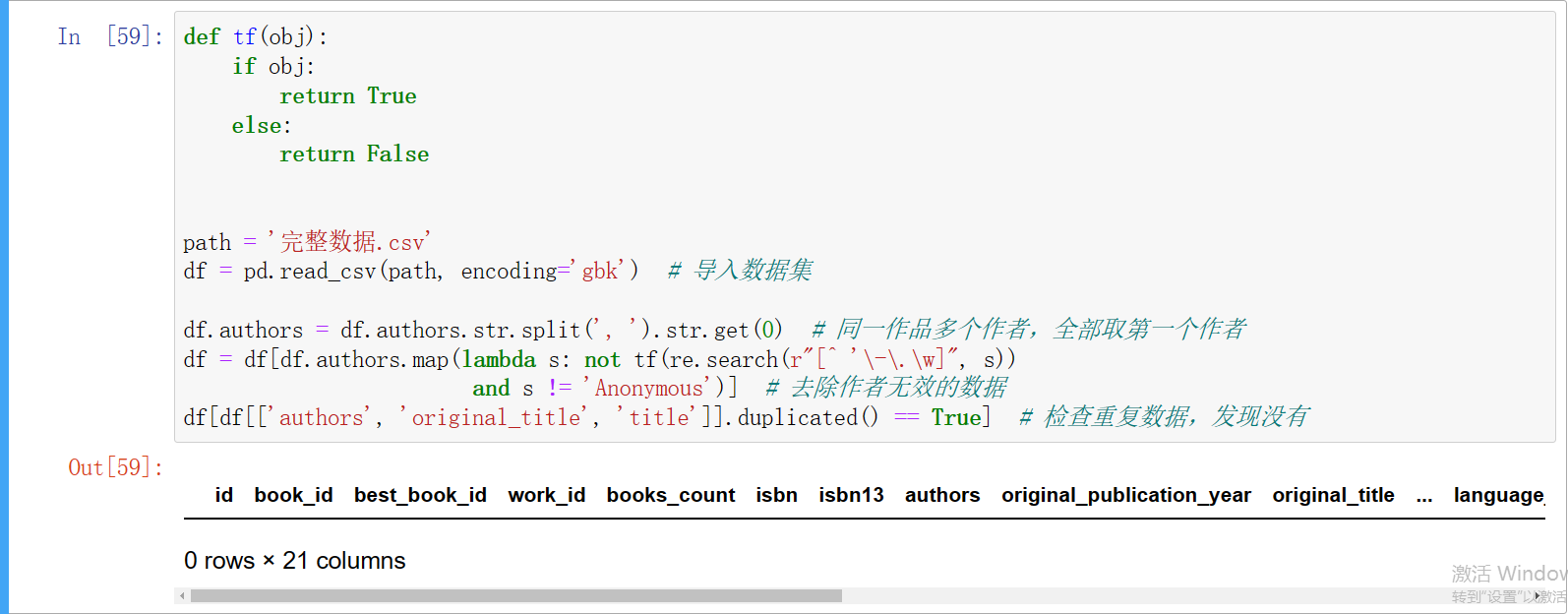
1. 各段位作者，不同书籍数量的作者人数分布情况；
2. 各段位作者的书籍评分标准差的分布情况。

# 研究过程

1. **引入模块**

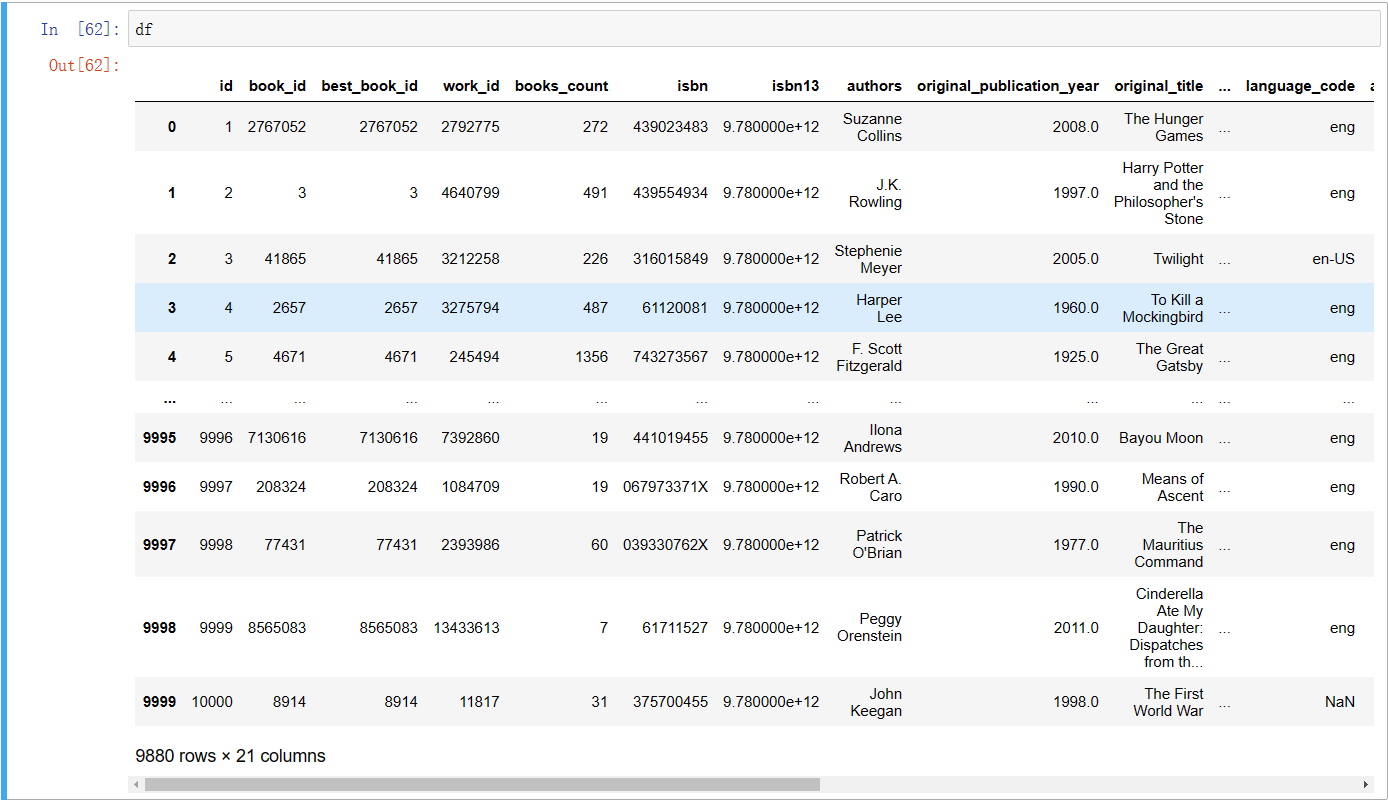


1. **数据清洗**



清洗后的DataFrame：

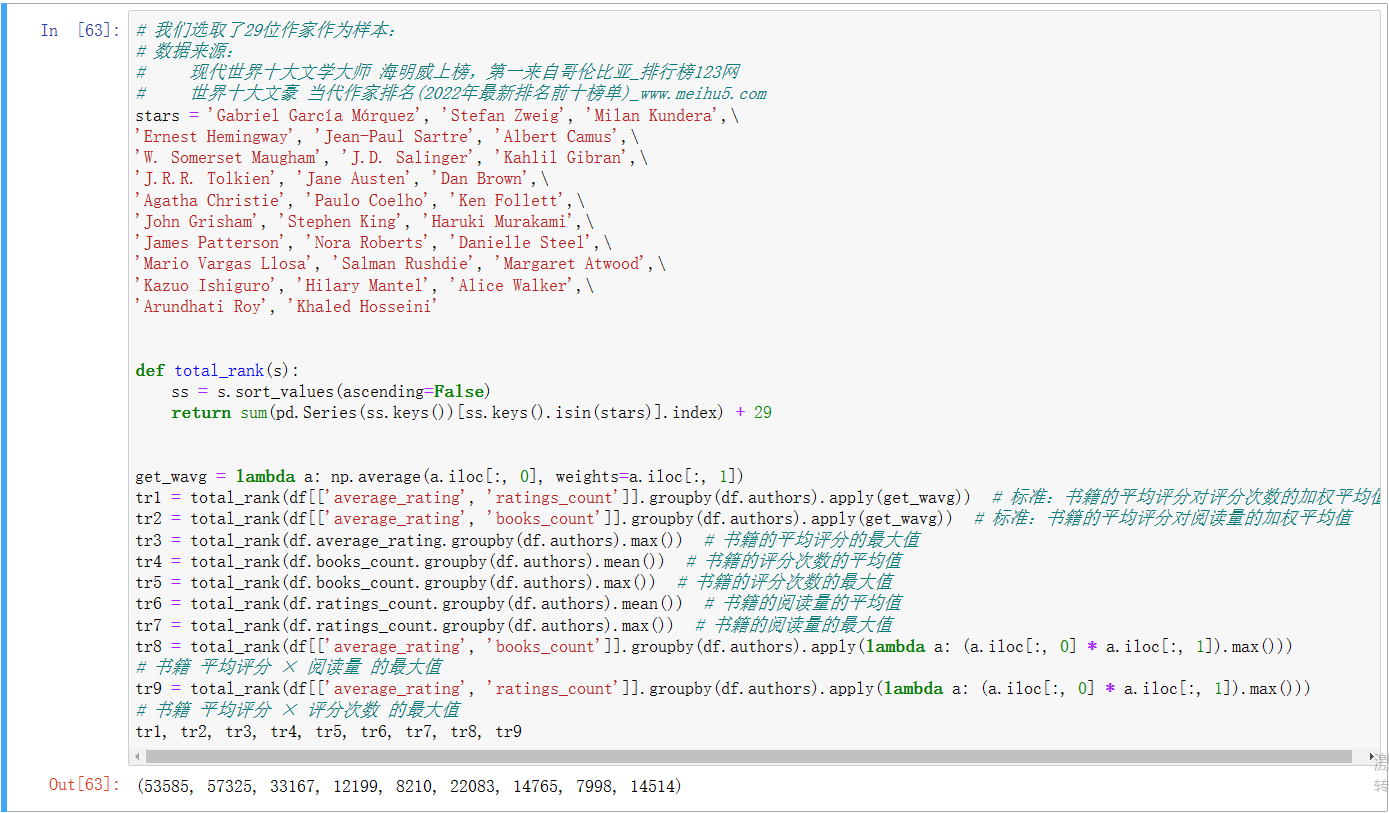
清洗后的“一万本好书的评分数据数据集”



1. **数据提取**
2. 寻找计量作者所获成就大小的最佳标准：

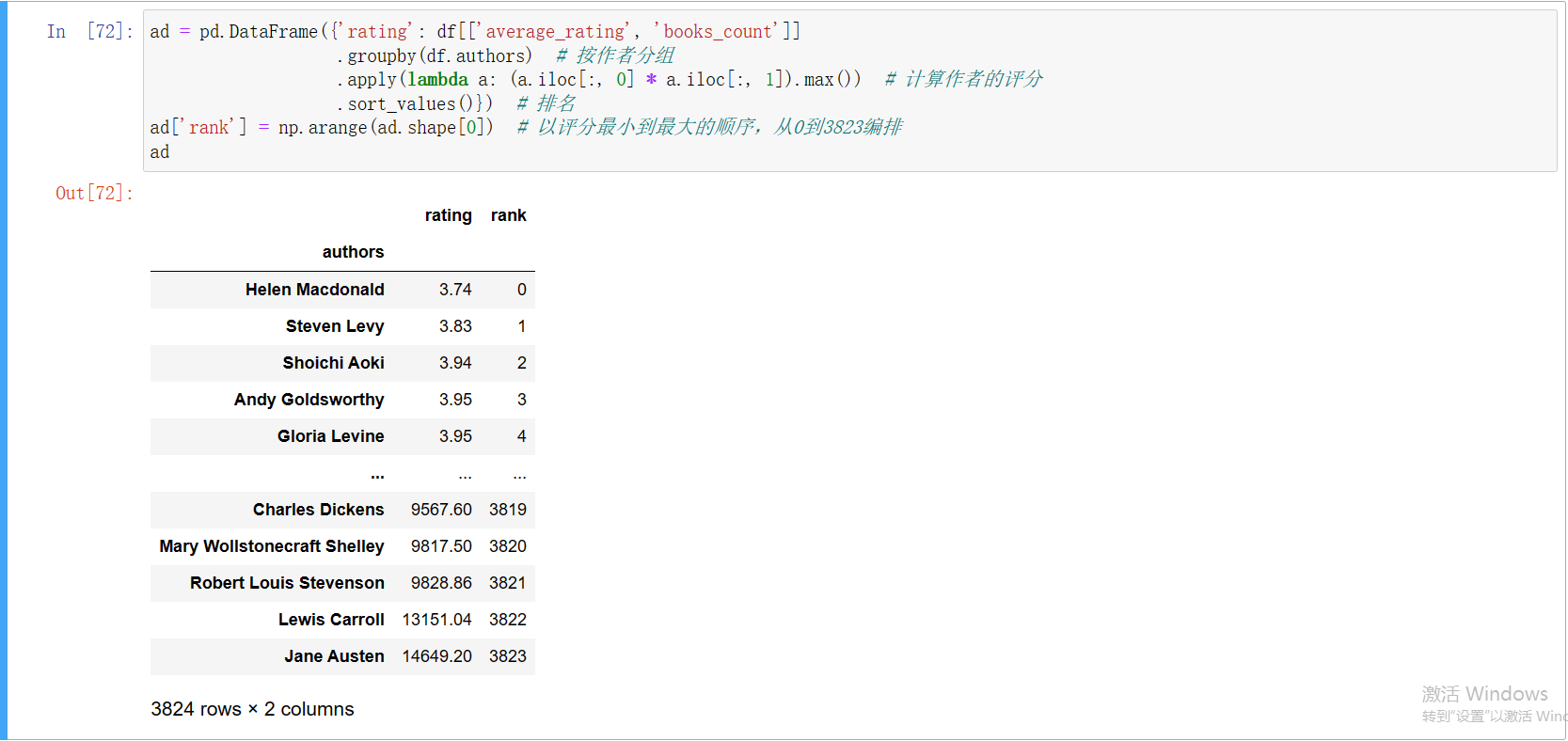
从网络上搜集29位最受认可的作家，换用不同的计量标准，统计在各计量标准下，他们的得分在数据表里的所有作者中的排名之总和，以总和最小的一个标准为最佳标准，作为之后研究的基础。

考虑到近代以前的作家，由于年代原因，可能对当代作家不具有代表性，研究意义较弱，因此本课题选取现当代作家来作为样本。

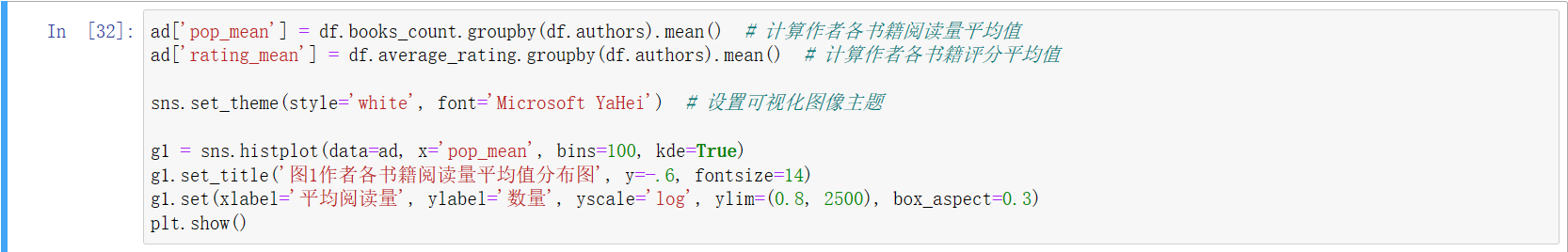


发现tr8是最小的，即**作者书籍的 平均评分 × 阅读量 的最大值是计量一个作者成就大小的最佳标准**。若认为书籍的平均评分能够代表读者对该书的普遍评价，则该标准就是作者的代表作品获得所有读者的评分总和，这就是该标准的意义所在。这也体现出一位作者的成就，不在于他写过作品的数量，也不在于他的作品的总体优劣，而在于他最好的少数几部作品的成就。这是符合本课题的生活常识的。

1. 计算所有作者的评分，生成一个新DataFrame，并排名：



1. **读者是否受到了对作者的从众心理的影响？**
   1. **设计和验证：**
2. 分别绘制作者各书籍阅读量平均值、评分平均值的分布图，体现出两类作者的区别。若读者不受从众心理的影响，即随机挑选书籍并完全客观根据书籍的优劣进行评分，则两个图像都应当接近正态分布。



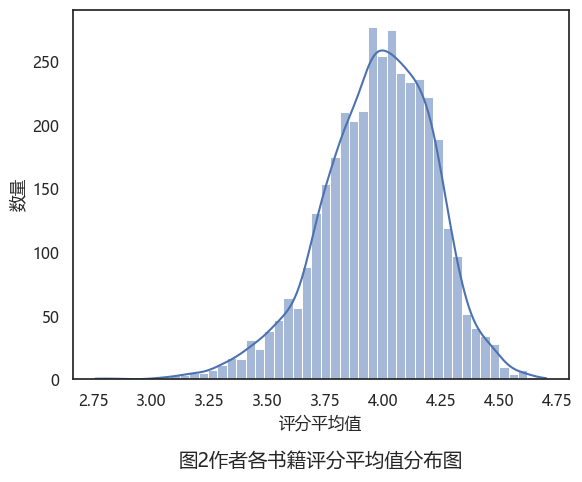
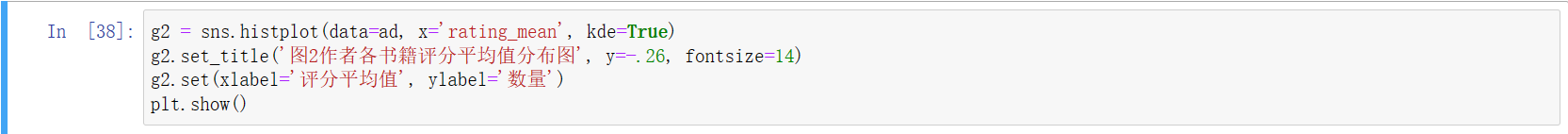
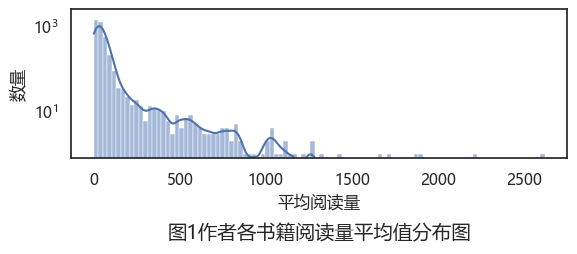
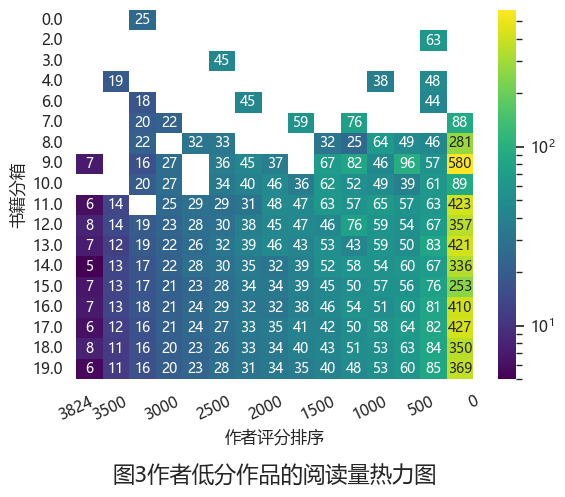


图1作者各书籍阅读量平均值分布图明显呈右偏态，说明读者会有意识地倾向于阅读著名作者的书籍。而图2作者各书籍评分平均值分布图偏态不明显，说明读者对作品的评分并不受作者名气的影响。

1. 为防止某些作者最著名的作品对阅读量造成数据扰动，影响结果，本课题对低分作品，控制书籍评分的变量，进行佐证。即将评分划为20档，作者等频分箱为15组，绘制阅读量的热力图。



如图3作者低分作品的阅读量热力图所示，越往右，颜色明显越浅，说明阅读量越高。这佐证了读者在挑选书籍时，受到了作者的名气的影响。

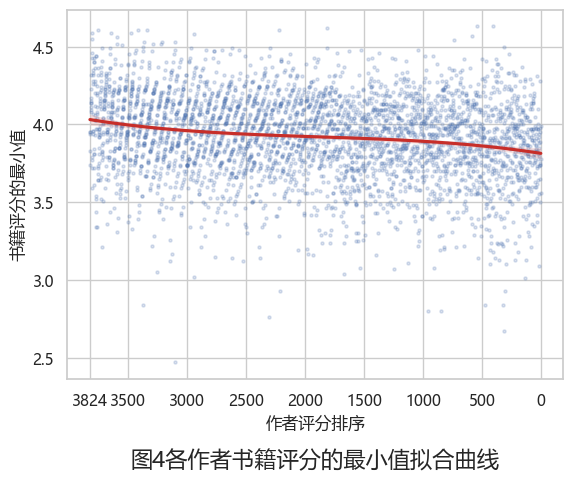
* 1. **结论：读者在挑选书籍时，受到了对作者的从众心理较大程度的影响；而在为书籍评分时则没有。**

1. **书籍评分与阅读量逻辑链**

**总问题：既然读者对书籍的评分是比较客观的，为何与书籍评分有关的计量作者成就大小的标准不够合理？书籍的评分和阅读量，哪一个对作者更为重要？**

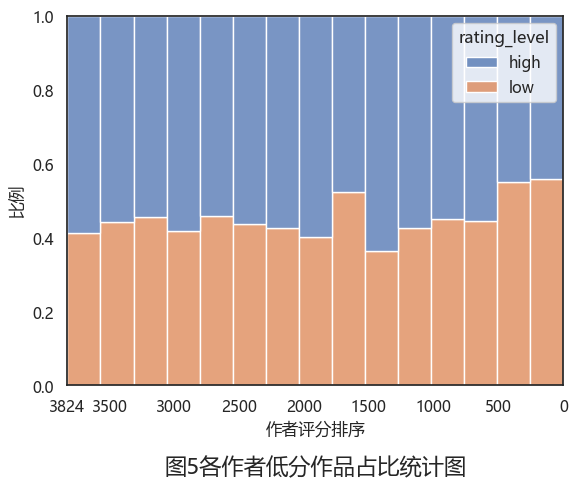
1. **为何评分的平均值不适宜作为计量作者成就大小的标准？**
2. 猜想：所有的作者，不论优劣，都写过一些好作品，但也写过不少差作品。
3. 设计：统计各作者写过的书籍的评分最小值，和低分作品的占比。若不论作者排名先后，评分最小值都相近，低分作品都有一定占比，则可知猜想成立。
4. 验证：

按作者排名，绘制各作者书籍的评分最小值散点图，并作出拟合曲线。



如图4各作者书籍评分的最小值拟合曲线所示，排名靠前的作者书籍的最低评分并不见得比靠后的作者高。

将作者等频分箱为15组，绘制各组中低分作品占比统计图。



如图5各作者低分作品占比统计图所示，几乎每组低分作品的占比都接近50%，最高分的作者的低分作品反而较多，超过一半。

1. 结论：说明**好的作者也并不一定是所有作品都优秀，他们也存在不少失败的经历。**大多数作者只有通过不断的练习和试错，才能创作出少数精品。

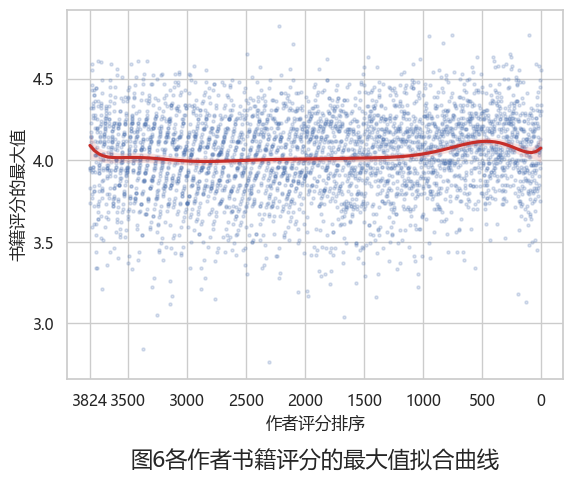
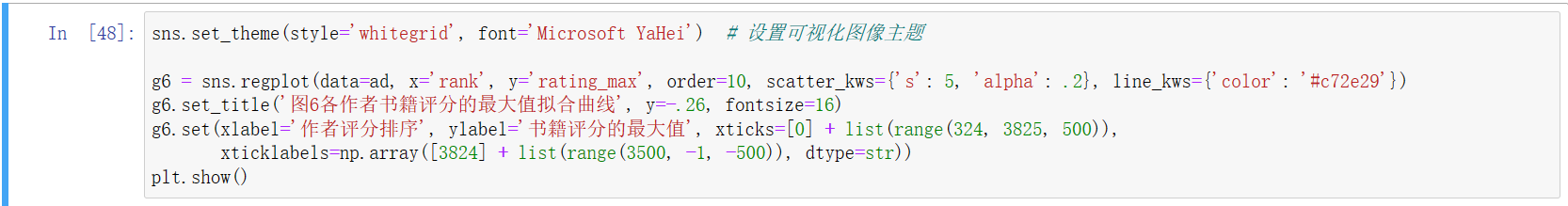
这也充分说明了书籍评分的平均值不能体现一位作者的水平，不宜作为衡量作者成就大小的标准。

1. **为何评分的最大值不适宜作为计量作者成就大小的标准？**
2. 数据现状分析



发现25%和75%的四分位数很相近，拉不开差距。

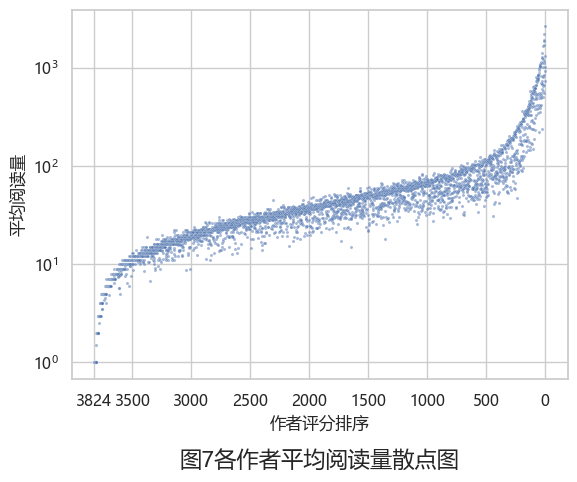
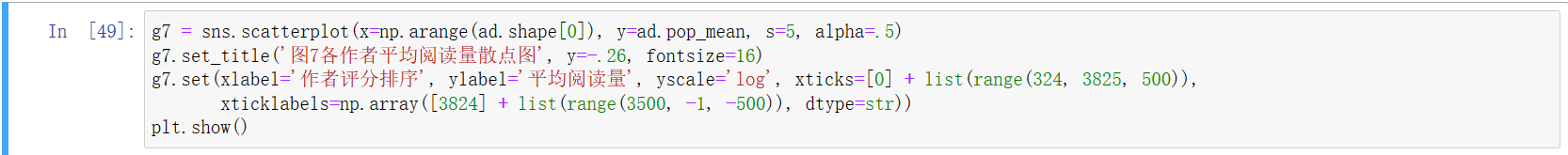
1. 按作者排名，绘制作者书籍评分的最大值的散点图，并作拟合曲线。



根据图6各作者书籍评分的最大值拟合曲线可见，发现评分的最大值也没有非常显著的规律，并且还拉不开差距。说明读者们对好作者的作品是褒贬不一的，作者的成就很可能也与书籍评分并没有很大关系。

1. **探究作者的成就与平均阅读量的关系。**

按作者的排名，绘制各作者平均阅读量的散点图：

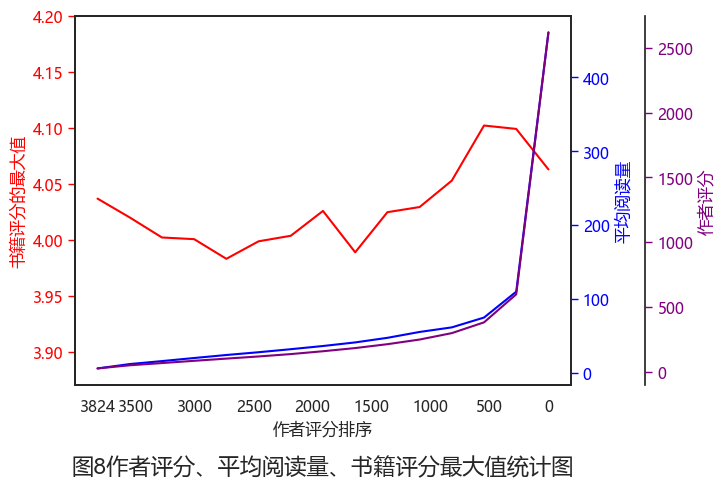


如图7各作者平均阅读量散点图所示，**越著名的作家阅读量越高，最著名的一些作家阅读量非常高**，这是毋庸置疑的。

另外，也基本可以看出，相比于书籍的评分，一个作者的成就与阅读量更有关系。

1. **书籍的评分和阅读量，哪一个对作者更为重要？**

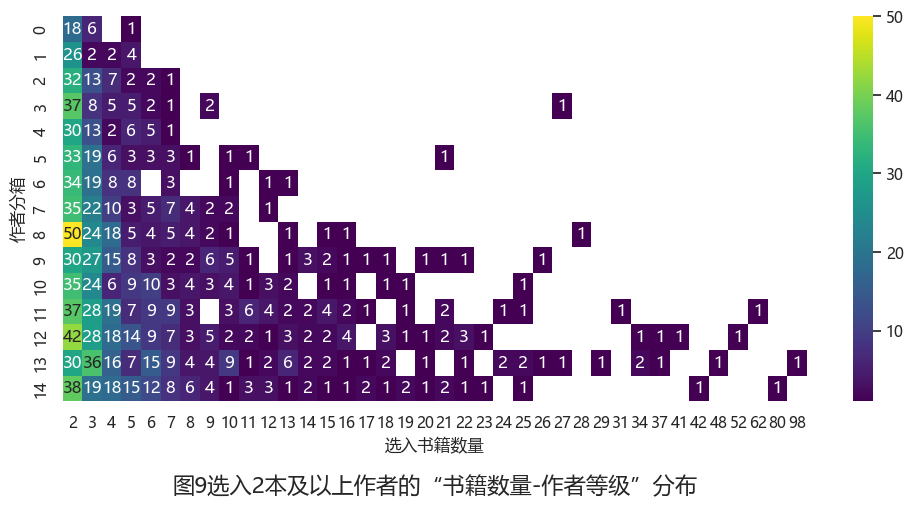
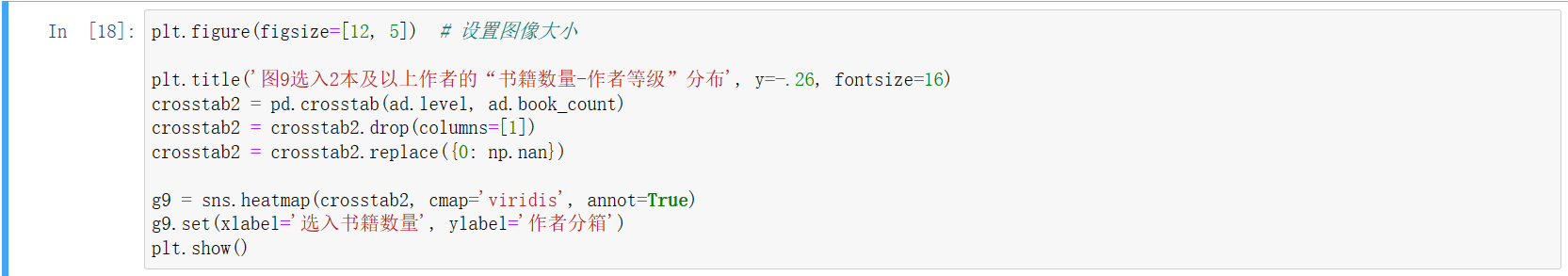
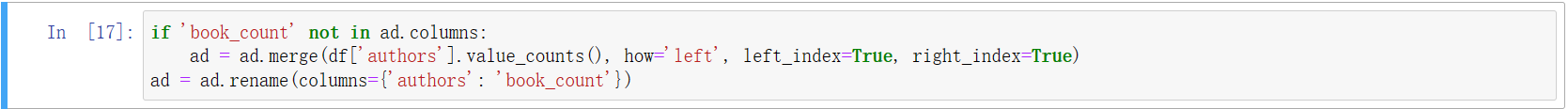
尝试将作者的评分、平均阅读量、书籍评分平均值三张图像作在一张图中：



如图8作者评分、平均阅读量、书籍评分最大值统计图，这进一步证明了**作者的成就还是与阅读量的关系更大，与读者的评分关系不大。作者需要写出能对广大群众有吸引力、有意义的作品。**

**总结论：与书籍评分有关的标准不适宜作为对作者所获成就的衡量，是因为一方面，各作者的书籍都有好有坏，评分平均值不能代表作者的真正水平；另一方面，相比于书籍的评分，阅读量对作者更为重要。**

1. **作者获得的成就与选入书籍的数量的关系**
   1. **猜想：**选入书籍数量越多者成就越大。
   2. **设计：**以选入书籍数量为横轴，作者15组的分箱为纵轴，绘制作者人数分布的热力图。由于仅选入一本书籍的作者数过多，影响图像绘制效果，没有必要，故舍去不作统计。
   3. **验证：**

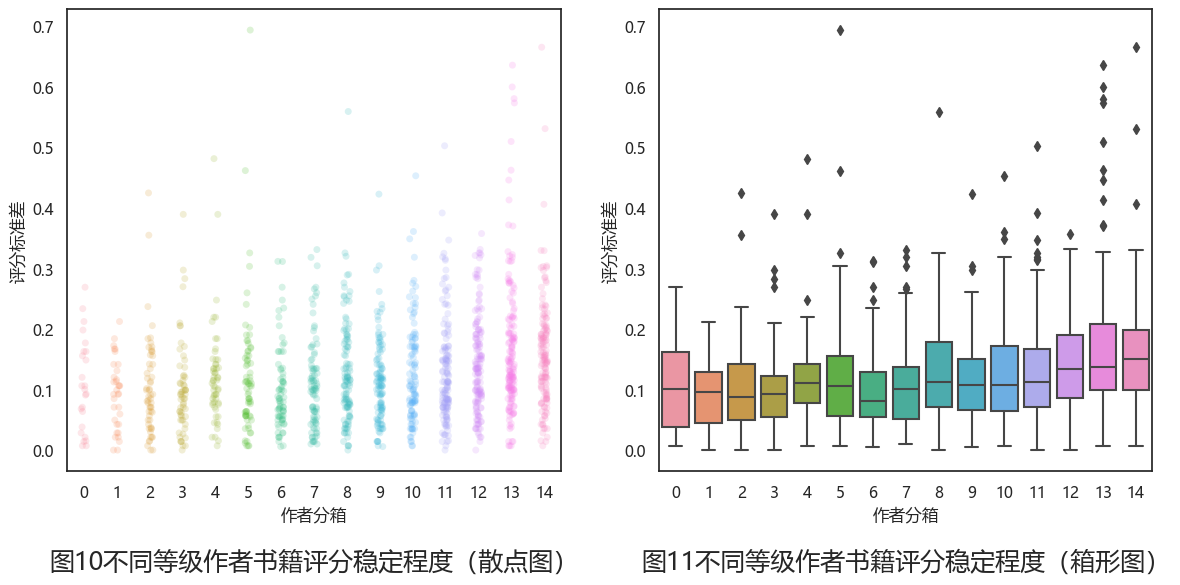
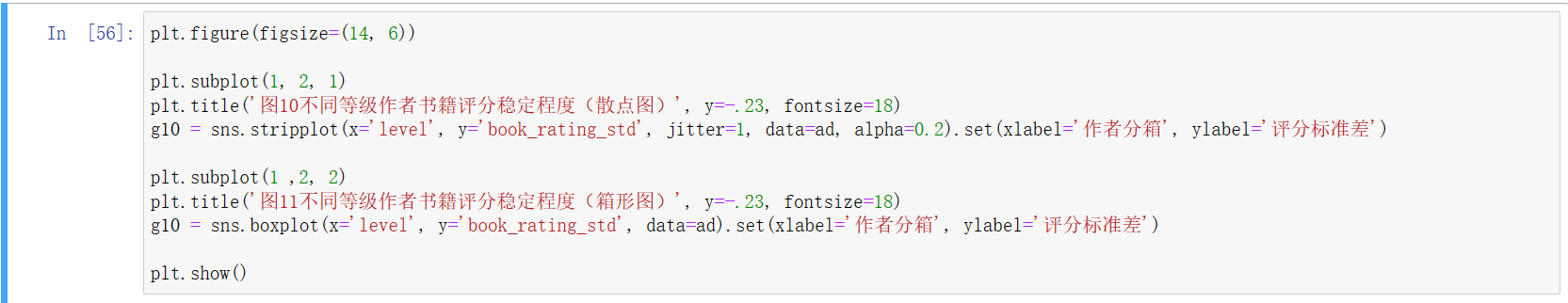


* 1. **结论：**如图9选入2本及以上作者的“书籍数量-作者等级”分布图所示，浅色多集中在左下部，说明**选入书籍数量较多者确实成就较高**。这也进一步佐证了有大成就的作者都是写作数量大且有不少好作品的人。

1. **作者获得的成就与水平发挥的稳定性的关系**

注：以作者各书籍评分的标准差为其水平发挥的稳定性。

1. **猜想：**成就越大的作者水平发挥越稳定。
2. **设计：**绘制各分箱中作者各书籍评分的标准差的散点图和箱线图。
3. **验证：**



1. **结论：**如图10不同等级作者书籍评分稳定程度（散点图）和图11不同等级作者书籍评分稳定程度（箱形图）所示，评分标准差总体较小，但右侧的作者水平发挥略不稳定于左侧。说明欲要成为成就较大的作者，不一定需要非常稳定的水准。

# **研究结论**

读者更倾向于挑选优秀的作家的书籍进行阅读，但对书籍进行评分时则都褒贬不一，几乎不受作者名气的影响，能够客观反映读者真实的内在想法。

好的作者要在大量的写作和试错中成长，在追求自己的个性、表达自己的想法、踊跃创新的同时，也要牢记大众是自己的写作对象，因此要写出能对广大群众有吸引力、有意义的作品。最终，也一定要出一二册精品佳作，方能取得较大的成就，受到专家和读者们的认可。

# **总结与展望**

本课题主要通过“一万本好书的评分数据集”，先确定优秀作者和书籍的评价标准并验证其准确性，然后研究了“读者是否受到了对作者的从众心理的影响”、“书籍评分与阅读量逻辑链”、“作者获得的成就与选入书籍的数量的关系”以及“作者获得的成就与水平发挥的稳定性的关系”，使用数据可视化的方法从多个角度剖析数据，从作者角度得出了获取较大成就所需要具备的一些条件。

本文的研究对象集中在作者身上，而数据集的收录限制在“一万本好书”之内，即使本文已经尝试验证对作者评价的准确性，但还是可能缺乏对作者本身总体性的认识，导致结果的偏差。希望能够进一步获取更多数据，修正研究的结果。