Steam-store-games数据分析报告

一、研究背景

电子游戏产业近年来呈现出蓬勃发展的态势，全球市场规模不断扩大，吸引了众多的开发者、发行商以及投资者。Steam作为全球最大的游戏平台之一，拥有海量的游戏以及庞大的用户群体，我选择kaggle平台上的steam-store-games数据集进行数据分析，试图通过结果来寻找影响steam平台发布的游戏的市场表现会受哪些因素的影响。

该数据集是由从Steam商店和SteamSpy API收集的数据构成，这个数据集提供了关于商店中游戏各方面信息，例如其类型和估计的拥有者数量。并且是在2019年5月前收集的，包含该日期前发布的商店中大多数游戏。未发布的标题也被移除，以及许多非游戏软件，尽管可能有些许遗漏。

二、研究问题

本次分析旨在探究影响steam游戏市场表现的因素，从单因素分析、相关性分析、建立预测模型，以及社会网络、文本、图像数据分析等方面推进，并提出了以下核心问题，在数据分析过程中逐个解决：

探索性数据分析(EDA)—业务现状诊断

1. 游戏发行生态诊断：分析 Steam 平台游戏发行年份分布、TOP30 开发商 / 发行商的作品产出节奏与类型偏好，定位 “发行高峰期特征” 与 “高产主体的核心竞争力”，为中小开发者选择合作发行商、规划发行时机提供参考。
2. 玩家拥有量分层：基于玩家拥有量（owners_est）的分箱结果，识别 “小众游戏（<5 万拥有量）、腰部游戏（5-100 万拥有量）、头部游戏（>100 万拥有量）” 的数量占比与标签共性，明确不同量级游戏的市场格局，为开发者判断目标用户规模、制定市场定位策略提供依据。
3. 平台市场价值评估：统计 Windows/macOS/Linux 三大平台的游戏覆盖量、价格区间分布及 “支持多平台游戏” 的占比，量化不同平台的 “用户渗透价值” 与 “商业化潜力”，为开发者决定平台适配优先级提供数据支撑。

预测性数据分析

1. 单因素业务关联分析：量化游戏核心指标（销量 / 玩家时长 / 威尔逊评分）与关键业务特征的关联性 —— “不同价格对销量的影响”（已有）“TOP20 标签（ Indie/Action/RPG）与玩家粘性的相关性”“成就数量是否**显著**提升玩家留存（中位游玩时长）”，为开发者制定定价策略、标签运营、内容设计提供量化依据。
2.
   - 目标变量（y）：游戏热度，可用 revenue、owners_est 或 stickiness 等指标。
   - 解释变量（X）：包括价格、平台、标签组合、开发商历史口碑、早期评价、成就数量、粘性、wlb_score等。

主要结论

🎮 Steam 游戏市场

1. **市场整体在快速增长**
   自2014年起，Steam平台上的游戏数量持续上升，说明平台生态活跃，对开发者来说仍是值得投入的渠道。
2. **爆款游戏极少，大多数是小众作品**
   - 约80%的游戏拥有者少于5万人（小众）；
   - 约18%属于“腰部”（5万–100万玩家）；
   - 仅有约2%是头部热门游戏（超过100万玩家）。
     这意味着“做出爆款”非常难，多数游戏只能服务特定小群体。
3. **低价更容易吸引玩家**
   大部分游戏定价在7美元以下。低价（甚至免费）能显著提升玩家数量，而高价游戏（>20美元）往往玩家很少——除非是知名大作。
4. **口碑和玩家规模正相关**
   玩家越多的游戏，好评率通常越高、越稳定。小众游戏口碑两极分化严重：要么极好，要么极差。
5. **价格与好评率弱相关，但不是因果关系**
   高价游戏平均评分略高，但这更多是因为它们筛选了愿意付费的核心用户，且品质匹配预期，而不是“贵=好”。
6. **成就系统对收入几乎没用**

设置大量成就并不能显著提升玩家粘性或收入。游戏好不好玩，核心还是内容和玩法设计。

7. **不同标签代表不同“赚钱逻辑”**
    - 单人剧情类游戏（如RPG、冒险）靠一次性高质量内容赚钱；
    - 多人联机、卡牌、免费游戏则靠长期互动留住用户；
    - “独立游戏”“免费游玩”虽然收入低，但可能有忠实小众群体，适合做细分市场。
8. **支持三个平台（Win+Mac+Linux）更赚钱**
   跨三平台发行的游戏，平均收入比只支持一个或两个平台的高出近一倍。**单纯从Windows扩展到双平台，收益提升不明显。**
9. **促销月不一定带来更多收入**
   数据显示，Steam的大型促销季（如夏促、冬促）并没有显著提高当月游戏的整体销售额。可能因为热门游戏不一定在促销月上线，或折扣拉低了单价。
10. **预测游戏热度很难，但好评率最关键**
    在所有因素中，**玩家好评率（wlb_score）** 是影响收入最强的指标；其次是定价。早期增长快或玩家粘性高，并不直接等于赚得多。

- 定价别太高，尤其对新团队；
- 重视玩家体验和口碑，比堆成就更重要；
- 如果资源允许，尽量支持三大平台，否则专攻单平台；
- 不必迷信促销期发布，内容质量才是长期立足之本。

三、研究过程

首先进行数据介绍：

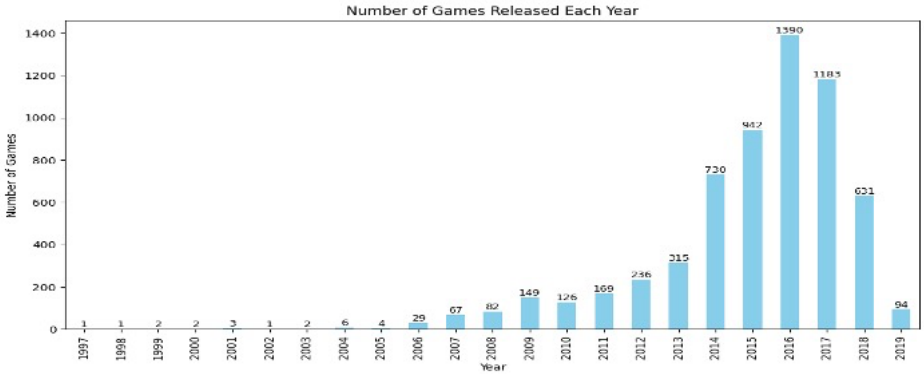
steam.csv:

字段名	含义说明
appid	每个游戏的唯一标识符
name	游戏名称
release_date	发行日期（YYYY-MM-DD）
english	是否支持英语（1 表示支持）
developper	开发商名称（多个以分号分隔）
publisher	发行商名称（多个以分号分隔）
platforms	支持的平台（windows / mac / linux，分号分隔）
required_age	PEGI UK 标准的最低年龄要求（0 多为未评级）
categories	游戏类别（如 single-player; multi-player）
genres	游戏类型（如 action; adventure）
steamspy_tags	SteamSpy 社区投票标签（分号分隔）
achivements	游戏内成就数量
positive_ratings	正面评价数（SteamSpy）
negative_ratings	负面评价数（SteamSpy）
average_playtime	平均游玩时长（SteamSpy）
median_playtime	中位数游玩时长（SteamSpy）
owners	估计拥有者区间（如 20000–50000）
price	当前全价（英镑 GBP）

- **缺失值**：数据集中存在少量缺失值，鉴于其数量较少（15行），我们选择直接删除包含缺失值的行，以保证数据质量。
- **重复值**：经检查，数据集中无完全重复的行。

游戏发行生态诊断

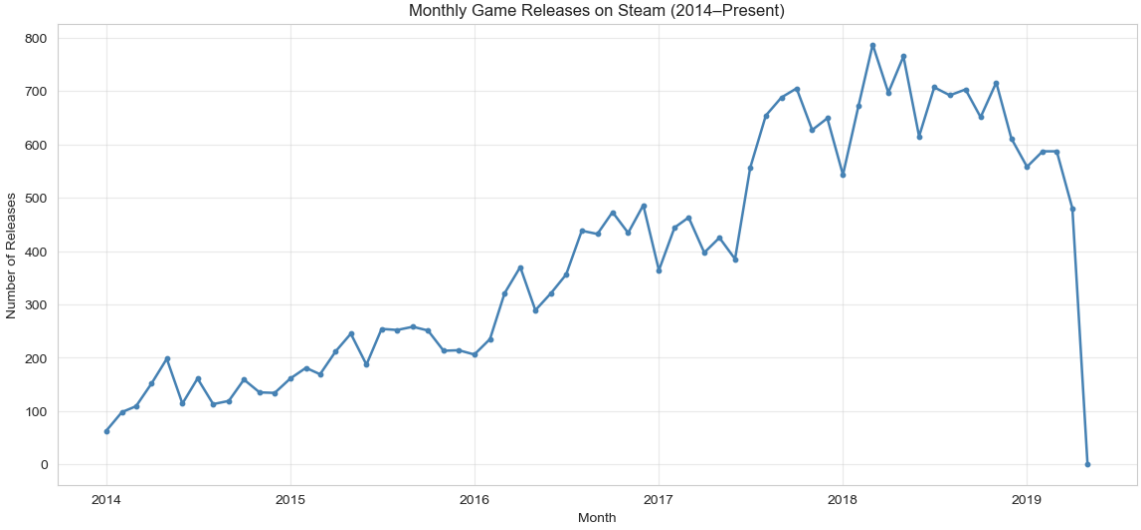
游戏发行年份/月份分布



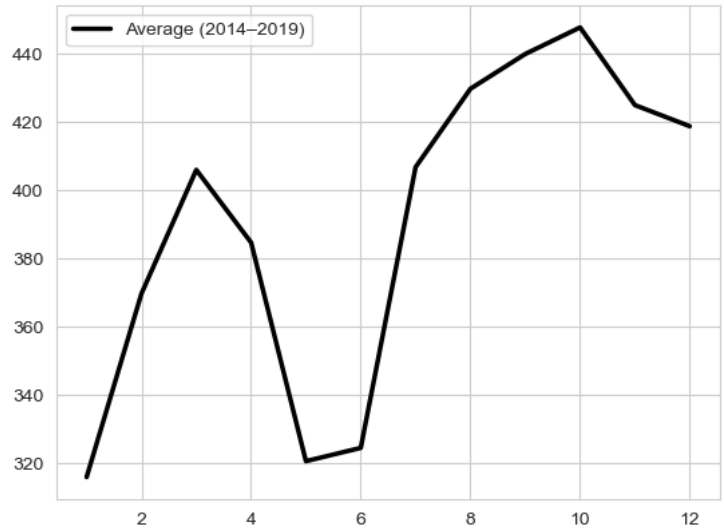
可以看到，数据时间范围从 1997-06-30 00:00:00 到 2019-05-01 00:00:00，可是2003年之前steam平台并未上线，经过调研发现这些游戏是在2003年前发布后兼容steam平台。

市场开始活跃的年份大约是： 2014。

为更细致地刻画 2014 年后市场的动态变化，将时间粒度细化至月份。月份序列表现出阶段性波动与周期性高峰，反映出平台在不同季节的发行节奏与市场活跃度。steam的促销季能大大刺激玩家的购买欲望，



接下来，按月份统计平均发行量：缺点在于可能会被个别年份同月极端值拉偏

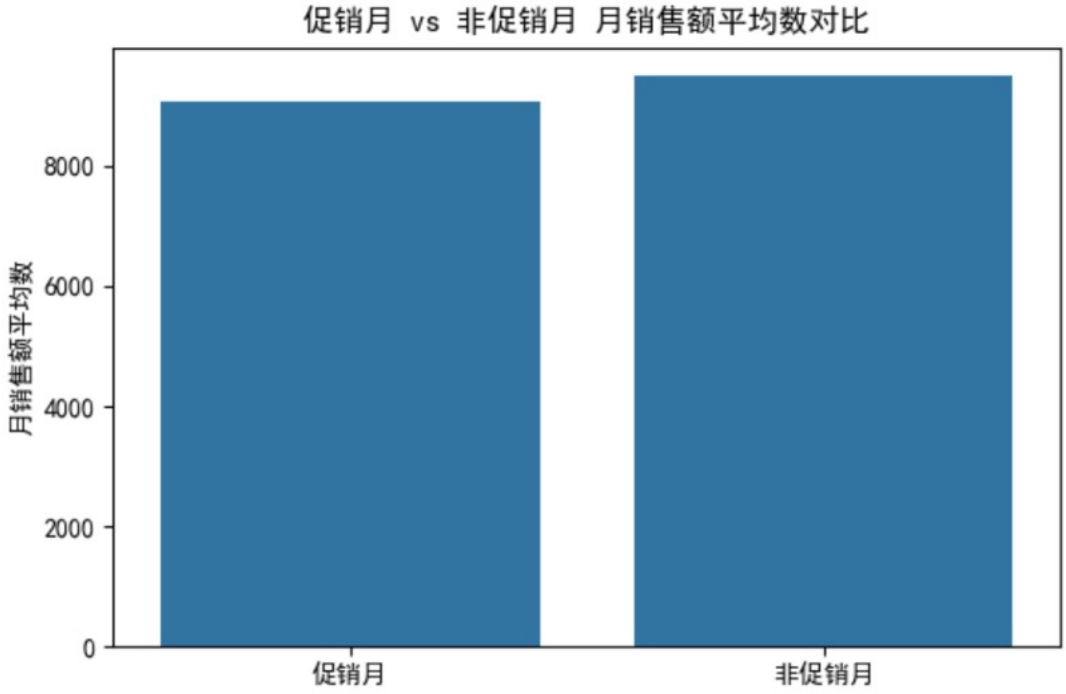


观察到：**发行高峰出现在3-4月，后半年整体发行量都较高。**

原假设 H_0：各月游戏发行数量的均值无显著差异（即不存在季节性）；

采用Kruskal-Wallis稳健检验，结果显示Kruskal-Wallis H = 2.925, p-value = 0.99168
未拒绝原假设：各月发行数量无显著差异（无明显季节性），由于汇总后样本较少，并无法完全排除这种差异的显著性。

根据steam促销习惯，将冬促、春促、夏促的对应月份（2,6,11,12）定为促销月份，其他月份定为非促销月份，分析促销与否是否对销售额有显著影响。



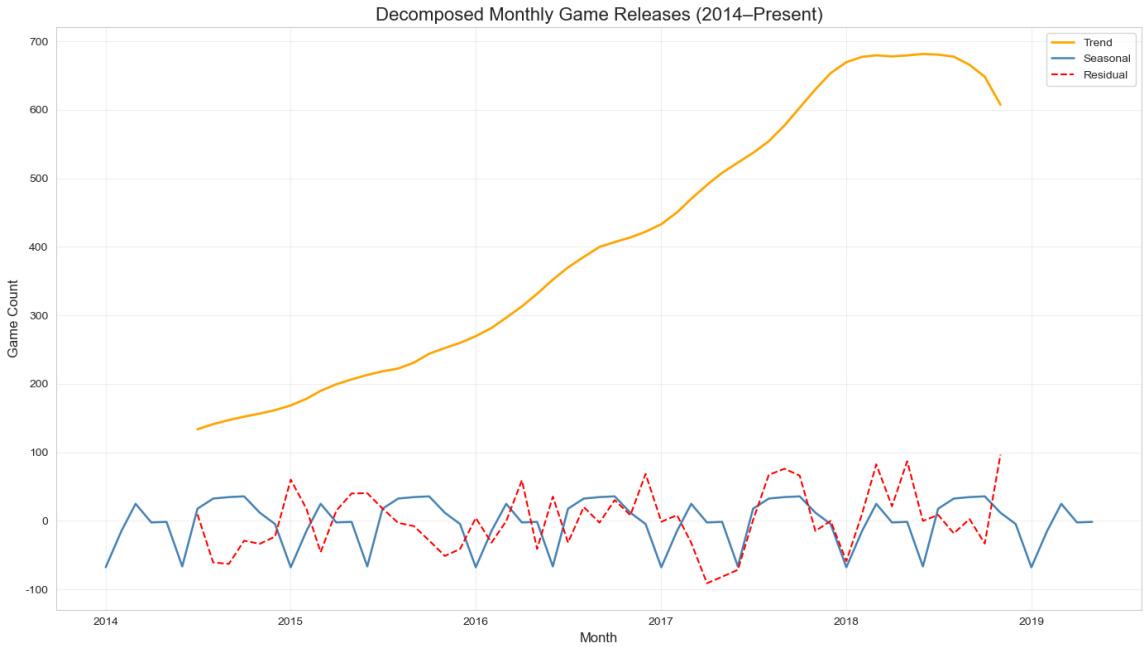
- 曼-惠特尼U检验：U=4003781.00, p值=0.3641
  - 曼-惠特尼U检验的p值显著大于0.05，我们没有足够的统计证据认为促销月与非促销月的游戏销售额显著差异
- t检验：t = 1.0153, p值 = 0.3101
  - t检验p值显著大于0.05，我们同样没有足够的统计证据认为促销月与非促销月的游戏销售额平均值有显著差异

首先，可能由于促销力度不足、玩家对促销不敏感或平台其他因素影响购买决策；

其次，也有可能是因为某些热门游戏的发行不在促销月，导致了非促销月销售额总体和促销月相差不大；

并且，月销售额也可能同时受游戏价格与购买人数的影响，可能因为游戏的打折导致销售额增长不大。

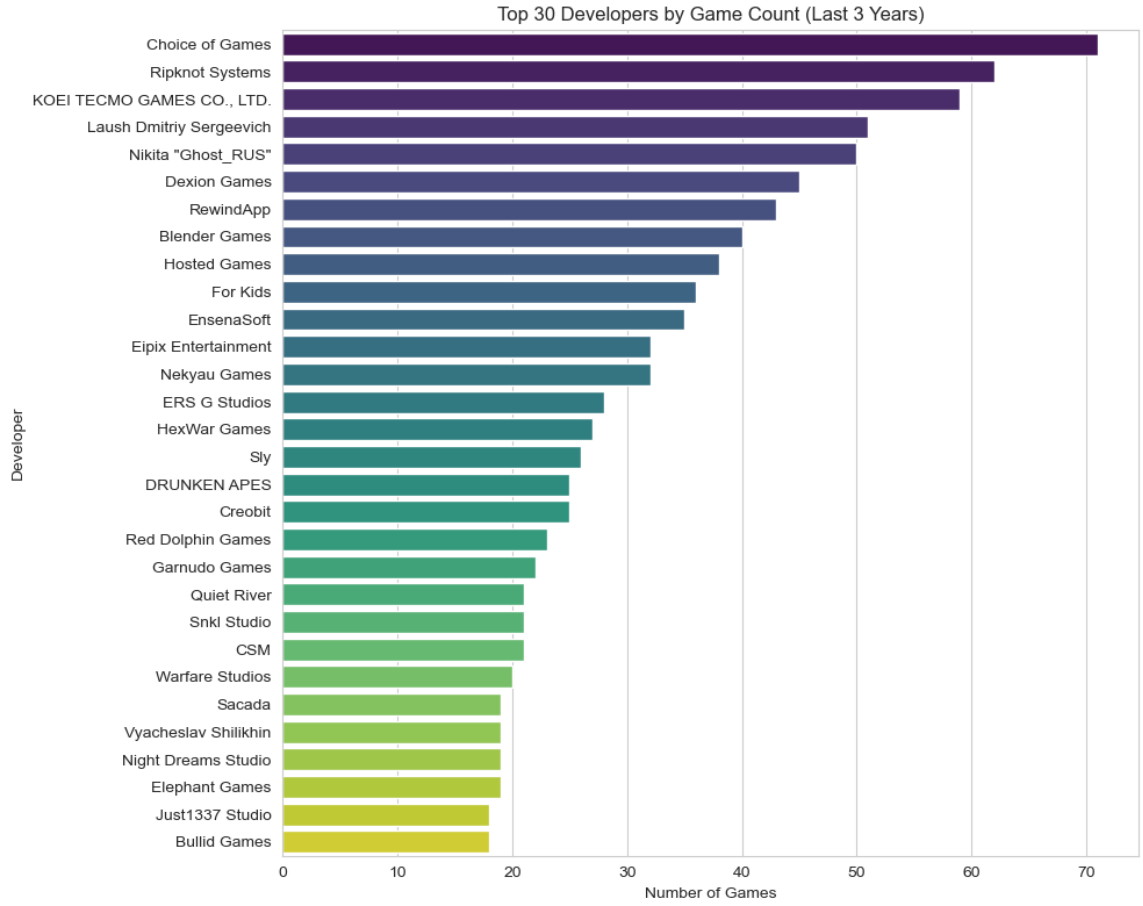
最后，滤除季节性波动，查看整体游戏发行量趋势：

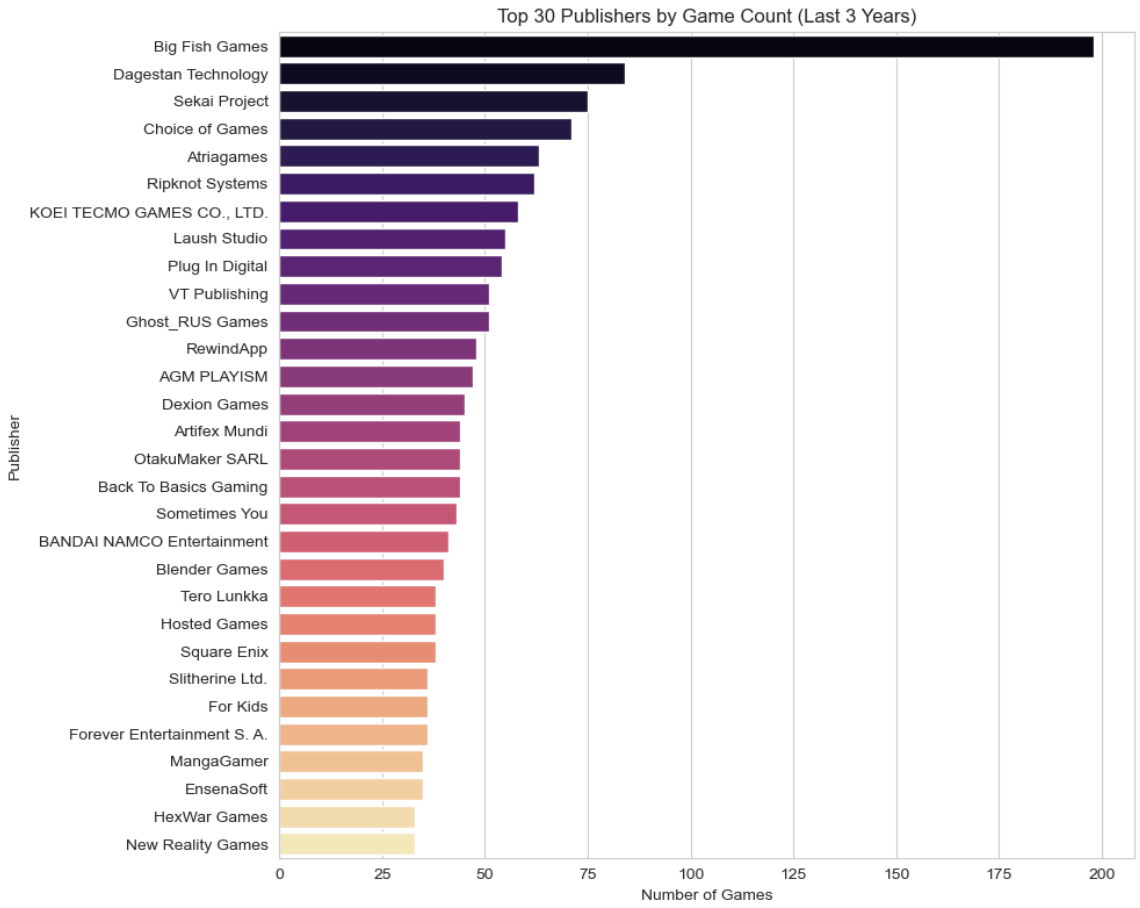


- 整体来看，Steam 平台的游戏发行量呈显著上升趋势，表明市场前景良好，为游戏公司选择该平台提供实证依据。
- 此外，时间序列表现出明显的短期周期性波动，提示发行商需合理安排发行节点，以降低直接竞争风险。

高产开发/发行商

近年高产开发商识别





前30名中既是开发商也是发行商的公司： For Kids HexWar Games Choice of Games Dexion Games Hosted Games Ripknot Systems EnsenaSoft KOEI TECMO GAMES CO., LTD. Blender Games RewindApp

作用主要有三点：

- 市场格局洞察：识别行业头部玩家及其主导力量，掌握竞争集中度和市场结构。

- 合作与竞品参考：为游戏公司在选择合作伙伴或评估竞品策略提供参考依据。

- 趋势与策略指导：通过高产者的类型、发行节奏和标签偏好，辅助制定产品定位、定价及发行时间策略。

⚠️ 数据局限性：由于缺乏这些开发商/发行商的更多背景信息及财务、市场数据，后两点的分析难以充分开展，仅能作为参考。

玩家拥有量分层

由于数据中每款游戏仅包含发行时的属性与总体统计指标，而未追踪其随时间的变化，本分析提供的是跨游戏平均意义上的洞察。即可以比较不同属性组合下游戏的平均玩家量、平均玩时及好评率，但无法推断单个游戏的发行后动态趋势。

1. 市场格局

```
- 小众、腰部、头部游戏的数量、玩家量占比（required age）
- 不同量级游戏的属性共性、差异
```

2. 策略指导

```
- 判断目标用户规模 → 决定做小众/腰部/头部
- 决定发行时间 → 避免竞争高峰期
- 确定价格策略、配置要求、标签定位
```

3. 趋势预测 / 风险提示

```
- 玩家量级结构变化趋势 → 市场容量、机会/风险
- 观察新兴标签或热门类型 → 潜在头部游戏方向
```

玩家拥有量分层分析指标体系

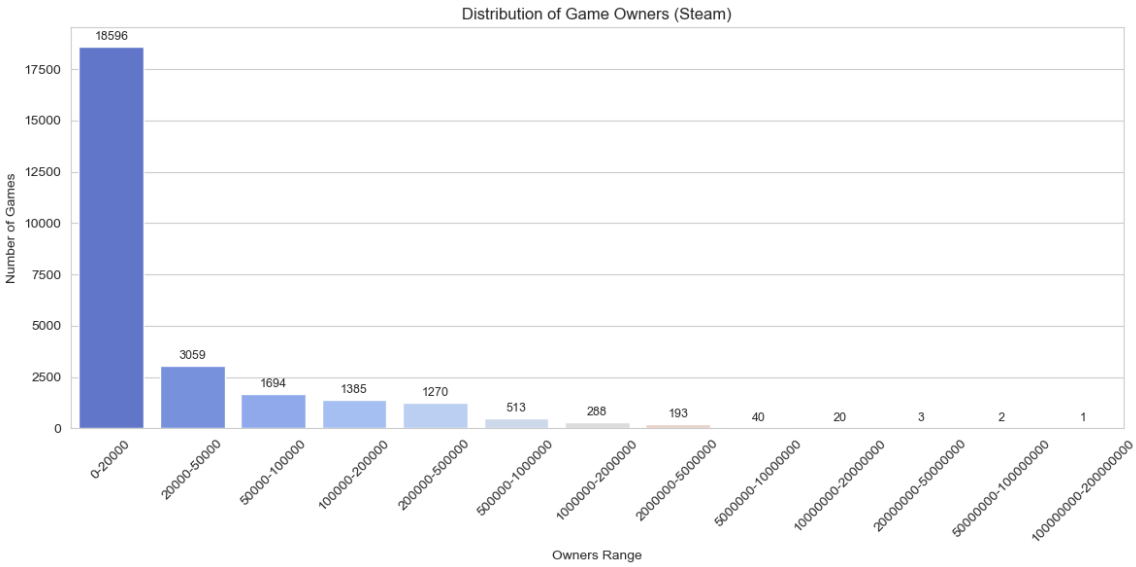
维度	指标	分层分析方法	目标
数量/玩家量	游戏数占比、玩家数占比	按小众/腰部/头部	市场结构洞察
标签	Steam 标签频率 Top-N	分层统计	产品定位参考
价格	均价/中位价	分层统计	价格策略
发行时间	年份、季度分布	分层统计	发布时间规划
评价	好评率、平均成就	分层统计	用户口碑分析
趋势	每类占比变化趋势	时间序列	市场预测 / 风险评估

玩家拥有量分层 × 属性矩阵

量级	游戏数占比	价格指标	发行时间分布	Steam 标签频率	用户评价	用户粘性
小众 (<5 万)	80.01%					
腰部 (5–100 万)	17.96%					
头部 (>100 万)	2.02%					

指标说明：

- **游戏数占比**：每类游戏数量 / 总游戏数
- **价格指标**：均价、中位价、折扣频次
- **发行时间分布**：按年、季度或月份分布
- **Steam 标签频率**：Top-N 标签占比
- **用户评价**：好评率、平均成就数
- **用户粘性**：playtime

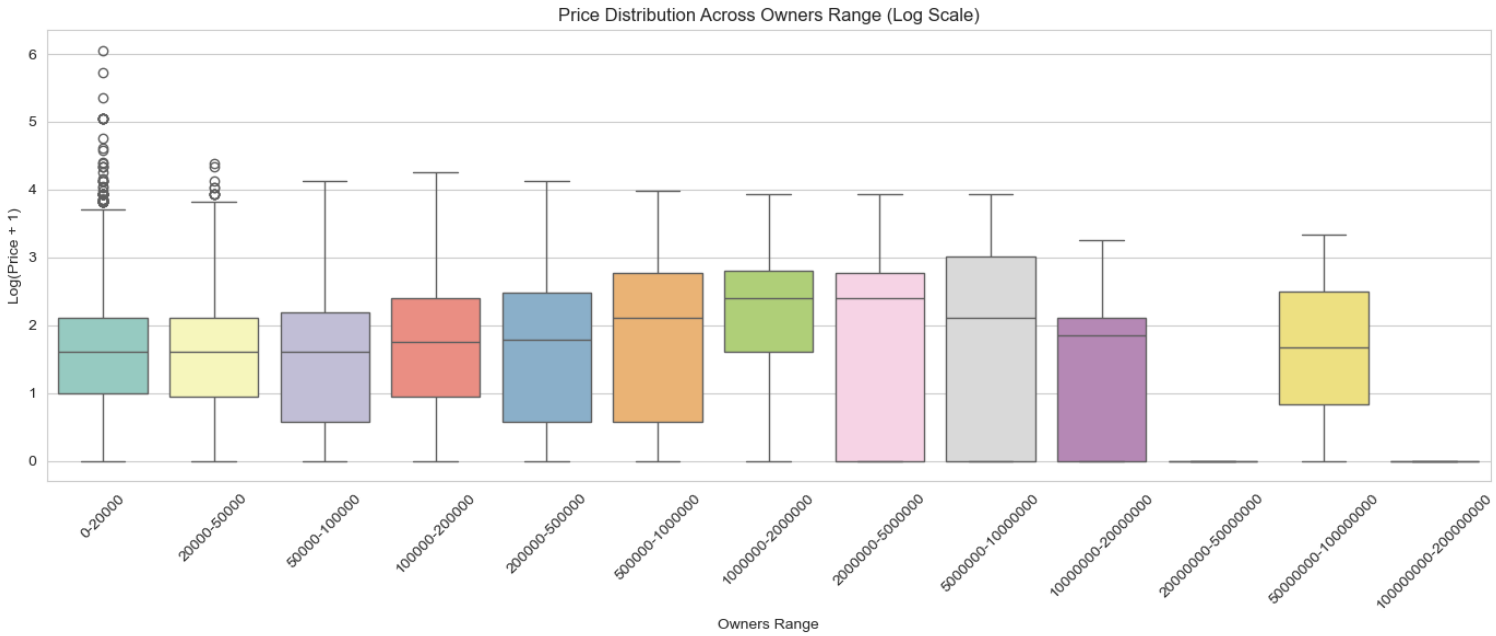


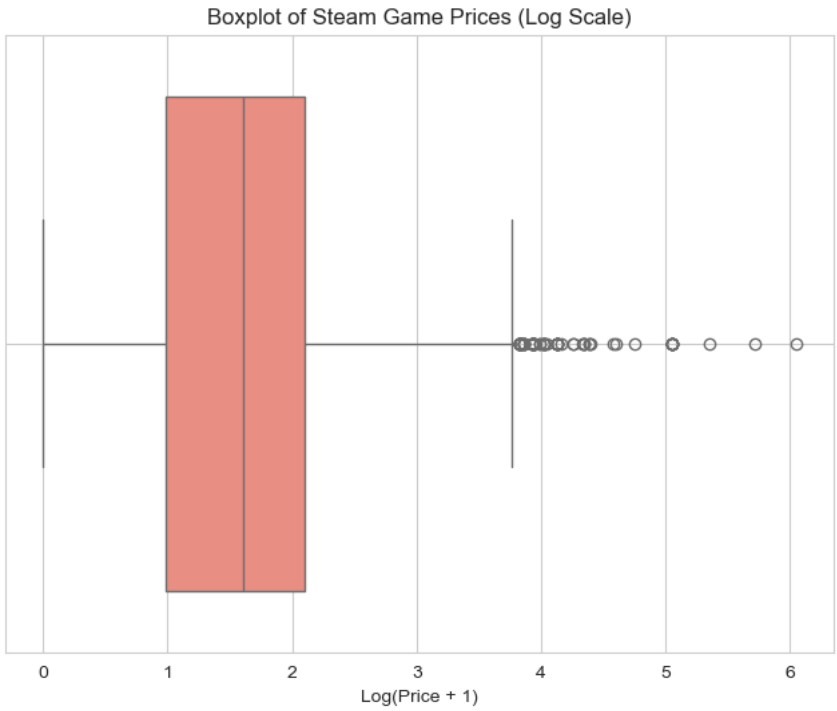
随着拥有者数目的增长，游戏数目不断降低。近八成的游戏拥有者数目处在[0, 50000]；

owners超过两千万的爆火游戏（Super-Head）：

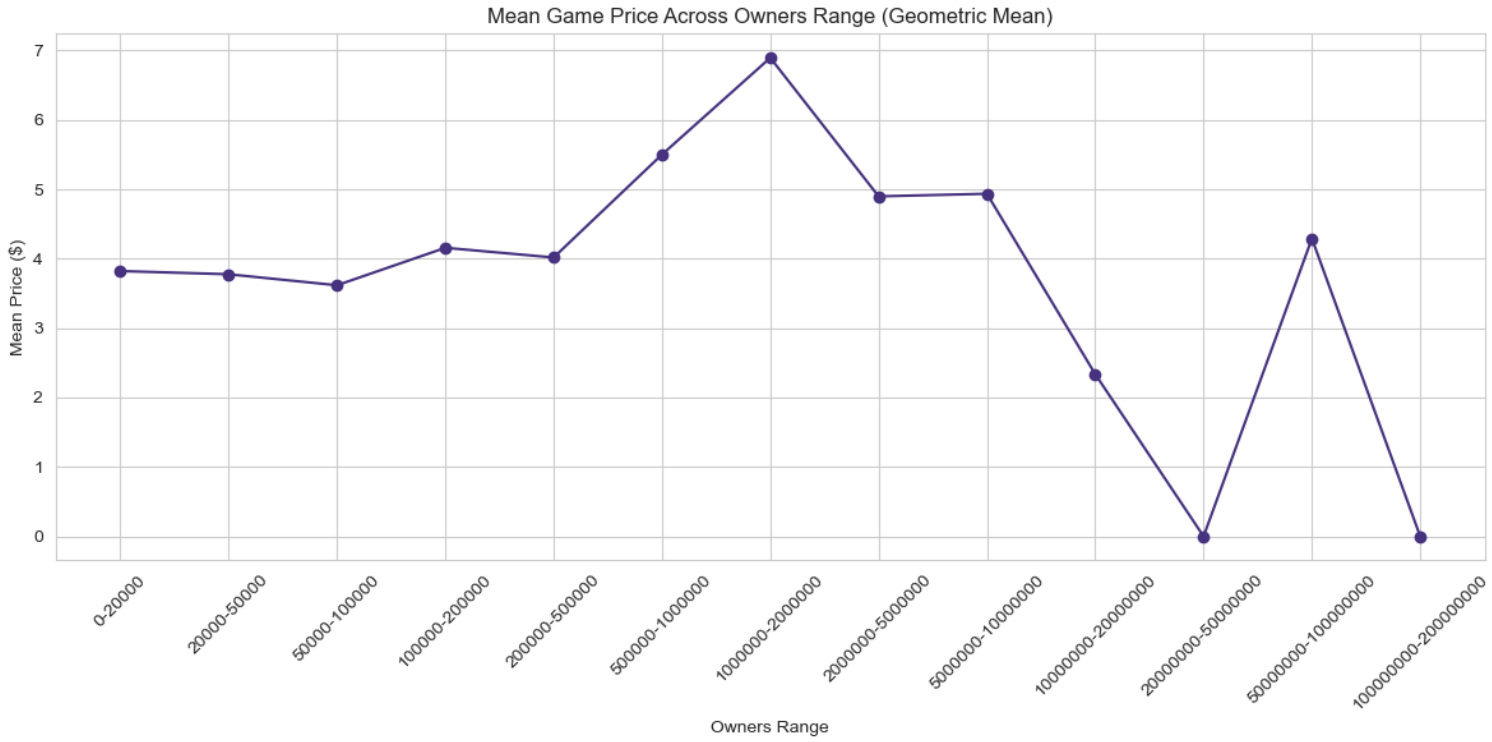
	name	owners	developer	publisher
12	Team Fortress 2	20000000-50000000	Valve	Valve
15	Dota 2	100000000-200000000	Valve	Valve
18	Counter-Strike: Global Offensive	50000000-100000000	Valve;Hidden Path Entertainment	Valve
1624	Warframe	20000000-50000000	Digital Extremes	Digital Extremes
3351	Unturned	20000000-50000000	Smartly Dressed Games	Smartly Dressed Games
12825	PLAYERUNKNOWN'S BATTLEGROUNDS	50000000-100000000	PUBG Corporation	PUBG Corporation

不同游戏的价格分布趋势：取对数（对数均值“低估”原始均值，解释为“价格的几何平均水平”，相比算术平均更稳健）





count	27064.000000
mean	6.078879
std	7.876284
min	0.000000
25%	1.690000
50%	3.990000
75%	7.190000
max	421.990000



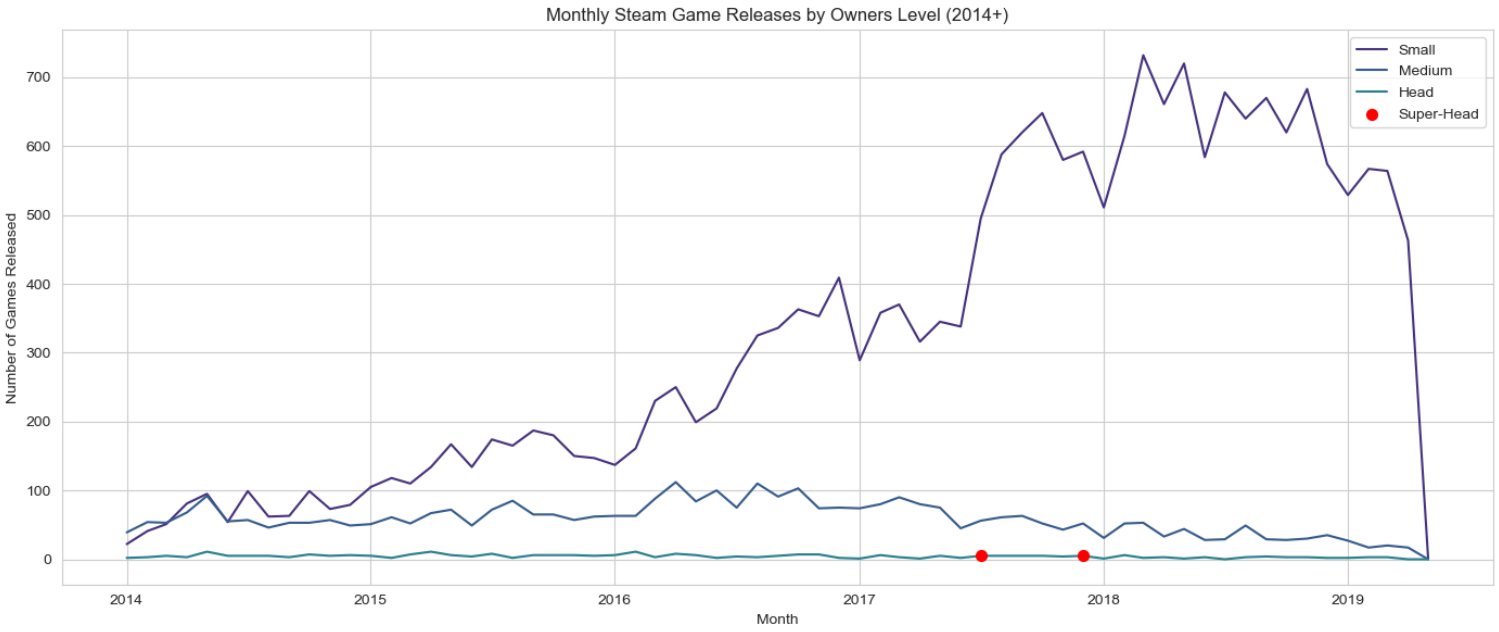
截至 2019 年，Steam 平台游戏价格分布呈明显偏低趋势：整体价格中位数为 3.99 美元，均值约 6.08 美元，大部分游戏价格集中在 1.69–7.19 美元区间（见箱线图）。分析玩家拥有量后发现：

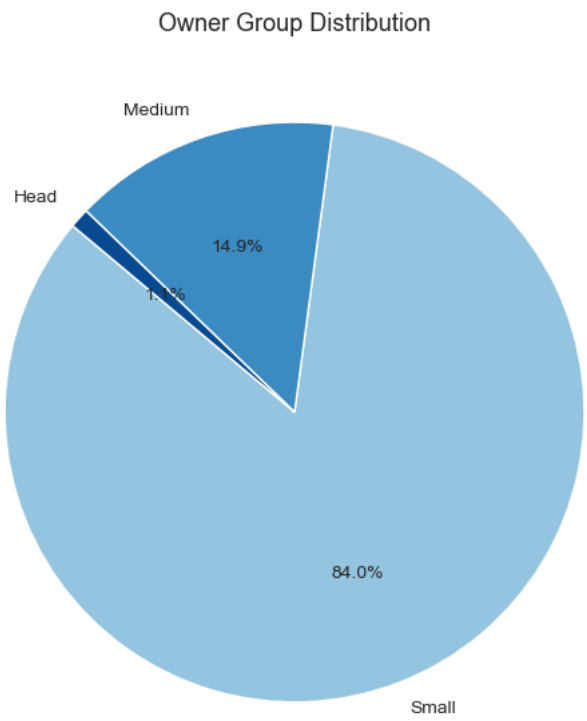
- **低价游戏（≤7 美元）** 平均玩家数显著高于高价游戏，甚至部分免费游戏拥有者数超过百万，显示低价策略有助于快速扩展用户群。
- **高价游戏（>20 美元）** 平均玩家数明显下降，部分玩家稀少的高价游戏可能并非质量不足，而是定价过高导致潜在玩家难以负担。
- 分层分析显示：小众游戏、中等玩家量游戏和头部游戏的价格分布存在差异，小众游戏价格偏高，腰部和头部游戏价格更贴近市场主流。

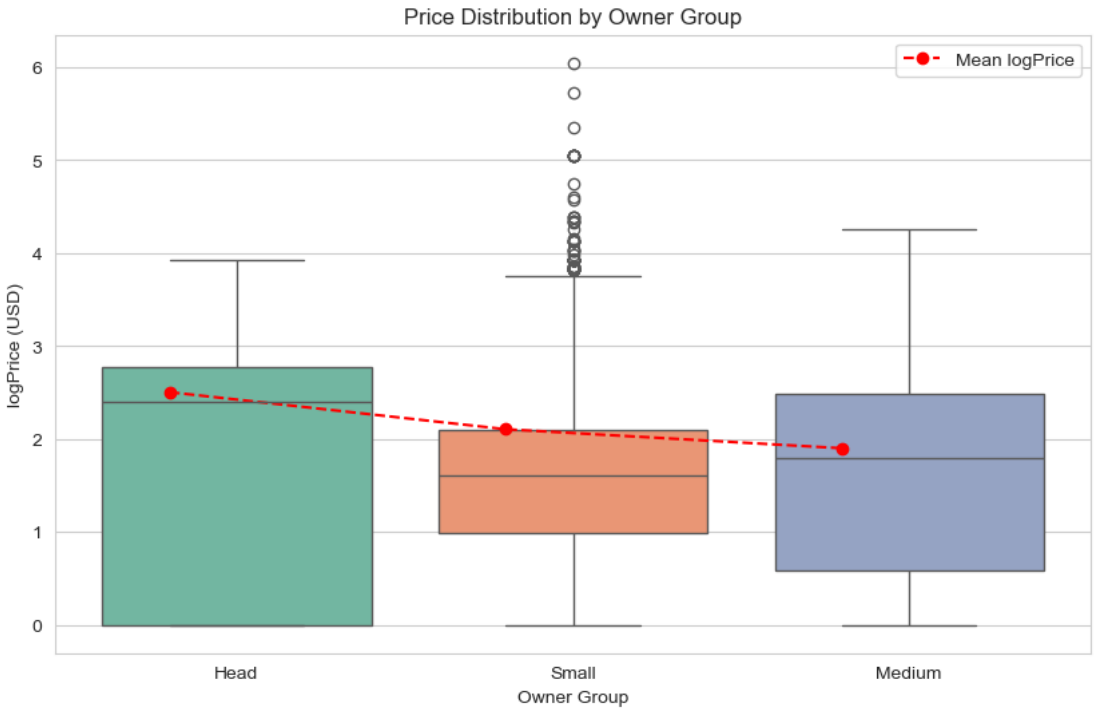
策略启示：

- 开发/发行者在平衡成本和收益的前提下，应考虑定价对用户规模的影响。合理的中低价格区间可帮助新游戏快速吸引玩家，提高市场渗透率。
- 对于小众高价游戏，可以通过折扣、促销或免费策略提高初期玩家量，从而提升口碑和长期收益。

分组发行量时间序列：







假设：头部（Head）、腰部（Medium）、小众（Small）游戏的价格分布没有显著差异。

说明：

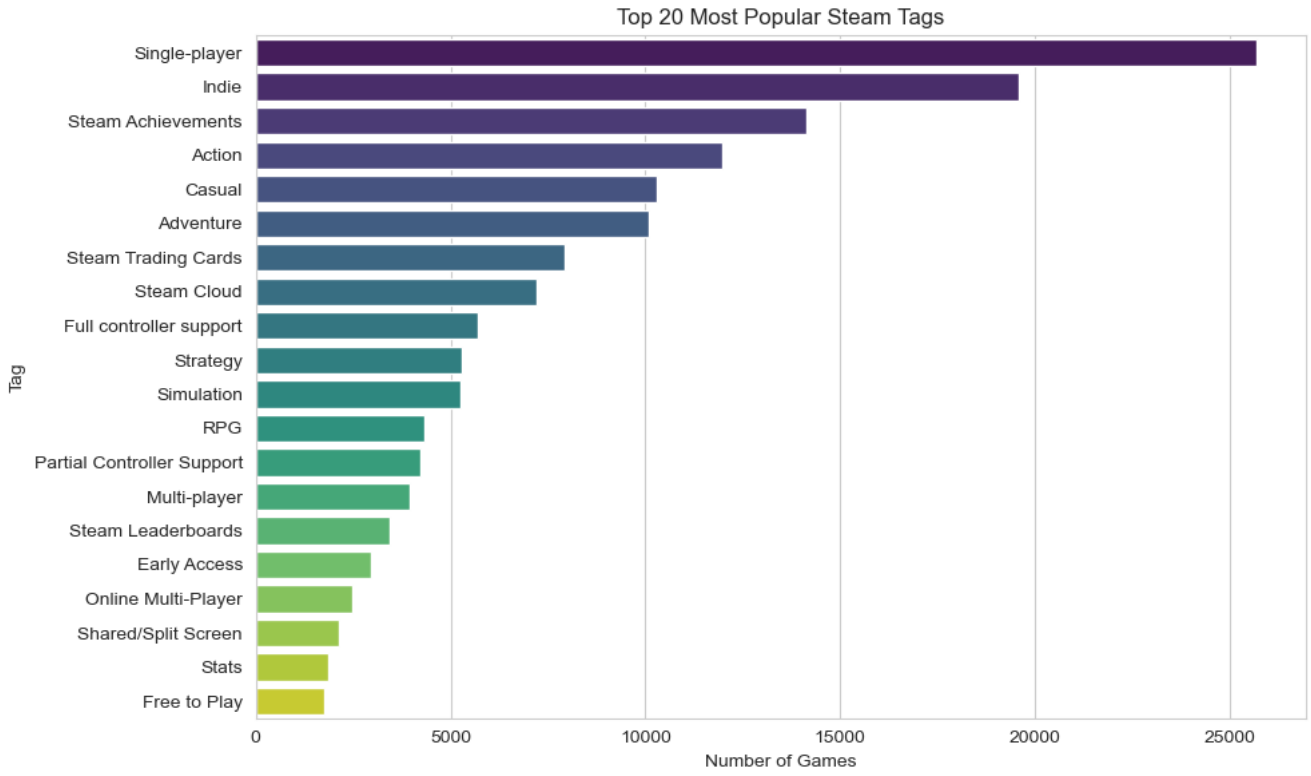
- 为保证每组样本量足够，本次分析未对极端价格值做严格界定，因此可能存在少量异常高价或低价游戏。
- 该假设关注的是总体均值或中位数的比较，而不是极端个例。
- 检验结果若不显著，并不意味着各类型游戏价格完全相同，只说明在样本统计范围内没有明显系统性差异。

Kruskal-Wallis H-statistic: 118.719, p-value: 0.000

结论：不同玩家量组游戏价格存在显著差异。

- 超级头部缺失：截至 2019 年 5 月 1 日，Steam 平台近 18 个月内未出现超级头部游戏（玩家拥有量 ≥ 2,000 万），显示顶级爆款游戏短期缺位。
- 腰部游戏趋势：与 2017 年及以前相比，中等玩家量游戏（腰部游戏）发行量呈下降趋势，尽管存在季节性小幅波动。
- 小众游戏趋势：小众游戏（玩家拥有量 < 5 万）发行量整体上升，同时保持显著季节性波动，高峰主要集中在春季和秋季，但最近 6 个月呈现下降。
- 头部游戏价格特征：头部游戏价格波动剧烈，多次出现极端值（如 2019 年部分月份价格突破 40 美元），整体均价显著高于中腰部和小众游戏，反映其定价策略灵活、溢价能力强，可能受大作发行和限时高价策略影响。
- 中腰部与小众游戏价格特征：价格走势相对平稳，长期均值保持在 5–10 美元区间，显示中腰部和小众游戏定价更趋保守，以稳定价格吸引目标玩家群体。
- 层级差异与市场效应：头部游戏与中腰部、小众游戏价格区间明显分层，体现 Steam 平台的“定价分层效应”——玩家规模越大，价格自由度与溢价空间越高；中小规模游戏则在价格上形成相对固定的竞争区间。

热门标签



分组评价情况以及用户粘性

大量游戏平均玩时和玩时中位数为0，应当剔除

Wilson score 公式如下（取置信度 95%，z=1.96）：

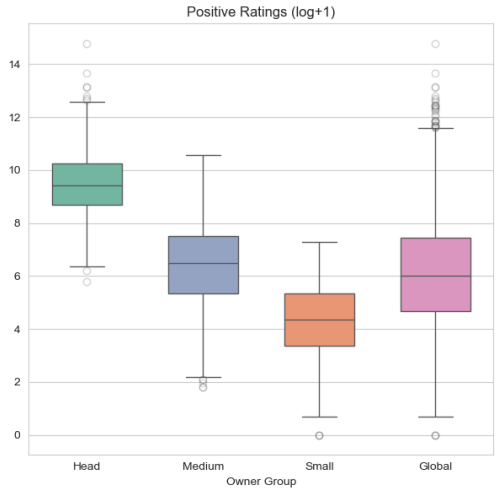
$$\text{Wilson Lower Bound} = \frac{\hat{p} + \frac{z^2}{2n} - z\sqrt{\frac{\hat{p}(1-\hat{p}) + \frac{z^2}{4n}}{n}}}{1 + \frac{z^2}{n}}$$

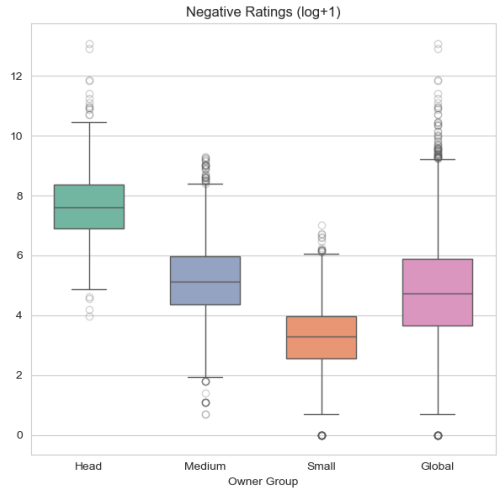
其中：

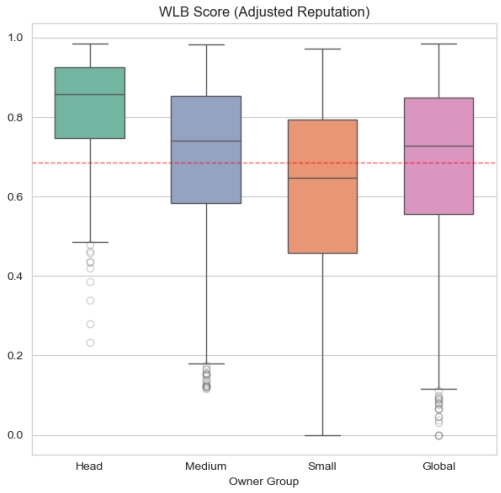
$\hat{p} = \frac{\text{positive}}{n}$,

$n = \text{positive} + \text{negative}$,

$z = 1.96$对应置信度95%







假设检验：不同玩家基数的组之间好评率wlb分数是否有显著差异（H_0：无显著差异）

📊 基本统计：

owner_group	count	mean	std
Head	545	0.822	0.129
Medium	3796	0.700	0.190
Small	1819	0.615	0.218

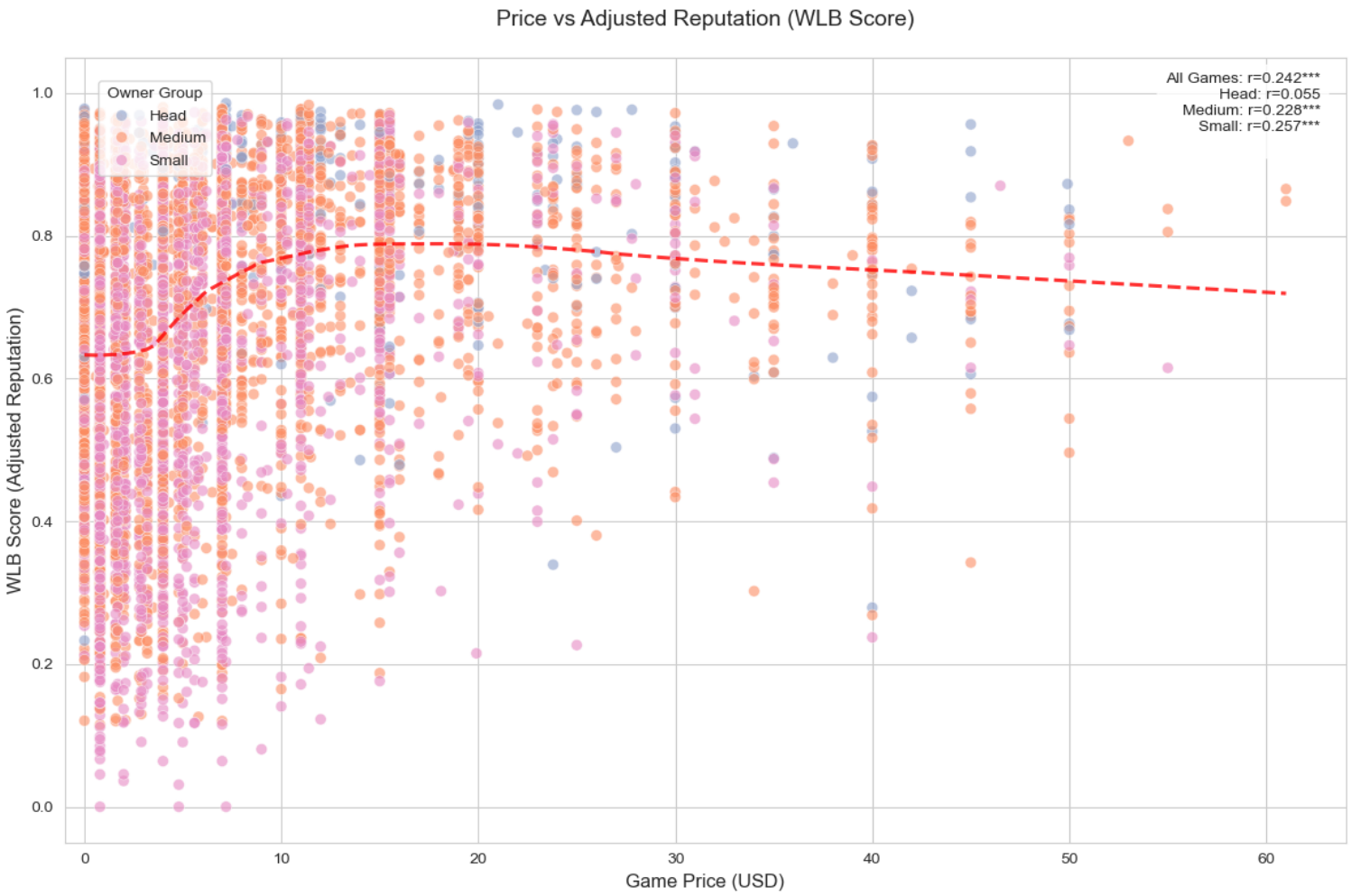
🔍 Kruskal-Wallis 检验结果: stat=500.439, p=0.000000
→ 显著差异

🎯 事后比较：
Small vs Medium: p_adj=0.000000 ✅
Small vs Head: p_adj=0.000000 ✅
Medium vs Head: p_adj=0.000000 ✅

Steam 游戏市场的口碑分布呈现明显的 “玩家基数分层效应”：玩家基数越大，口碑（经修正后）越好、越稳定。这种分层本质是 “品质筛选 + 规模效应 + 玩家分化” 共同作用的结果 —— 头部游戏靠品质和规模形成口碑壁垒，腰部游戏在竞争中寻求稳定，小众游戏则因品质参差不齐和样本量偏差导致口碑两极分化。

整体而言，**玩家基数提升有助于提升好评率。**

游戏价格与评分的相关性

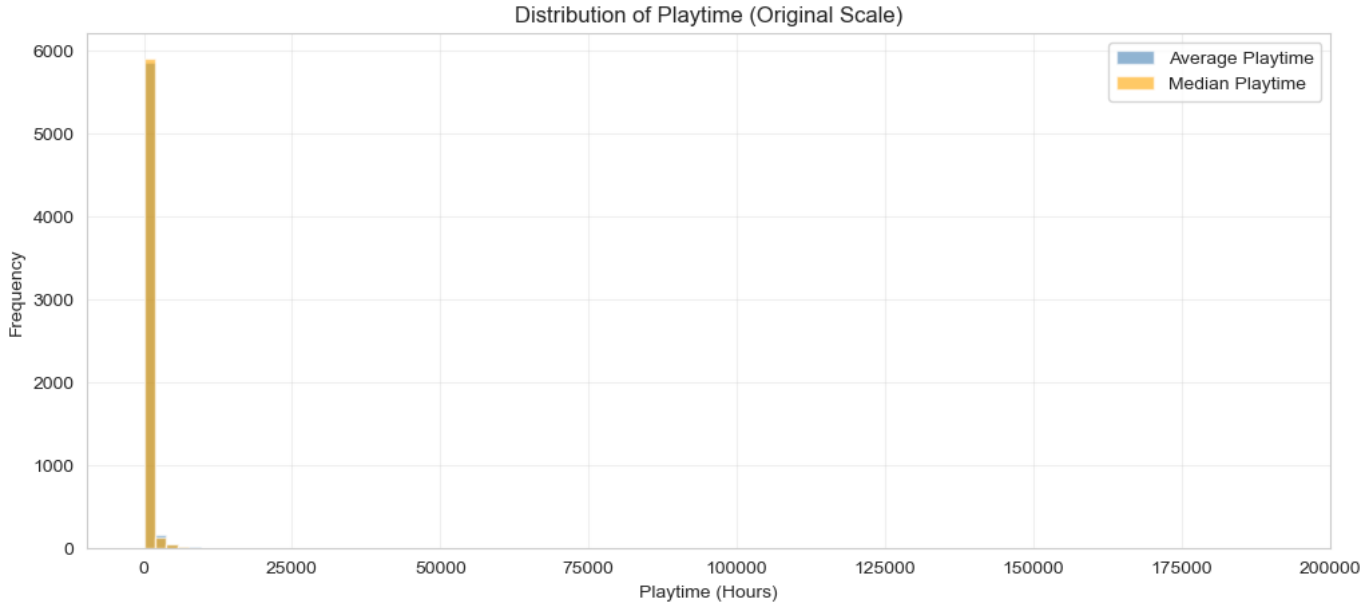


- 散点图整体较为分散，且数据集中于小价格区间，高价游戏样本占比低
- 整体来看，游戏价格与调整后声誉呈显著弱正相关，不同层级游戏关联强度存在差异：小众游戏（Small: r=0.268*）> 腰部游戏（Medium: r=0.243**）> 头部游戏（Head: r=0.203），均通过统计显著性检验（p<0.001）。
- 高价游戏的高评价并非源于价格本身，核心是 “价格筛选目标用户 + 品质匹配预期 + 策略兑现价值” 三者协同作用的结果。
- 分层特征表明：小众游戏价格与声誉联动性最强，头部游戏因品牌 / 生态壁垒，价格对声誉的直接影响相对最弱。

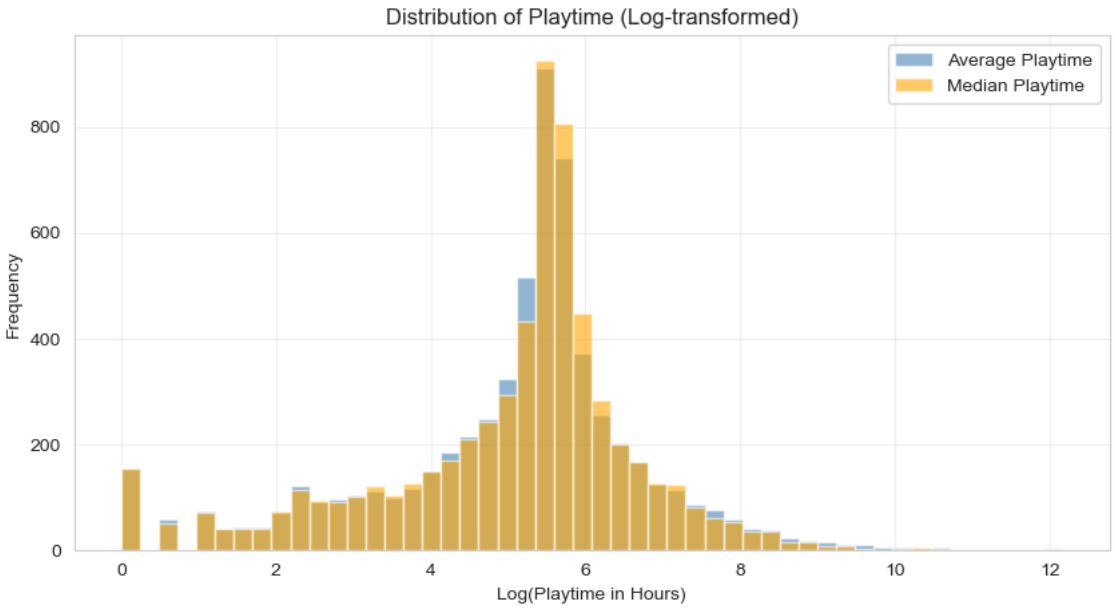
以上关系在小价格区间（20美元以下较为可信）

高价游戏的高评价并非一定是 “价格本身带来好评”，更可能是 **“价格筛选用户 + 品质匹配预期 + 策略兑现价值”** 三者共同作用的结果.

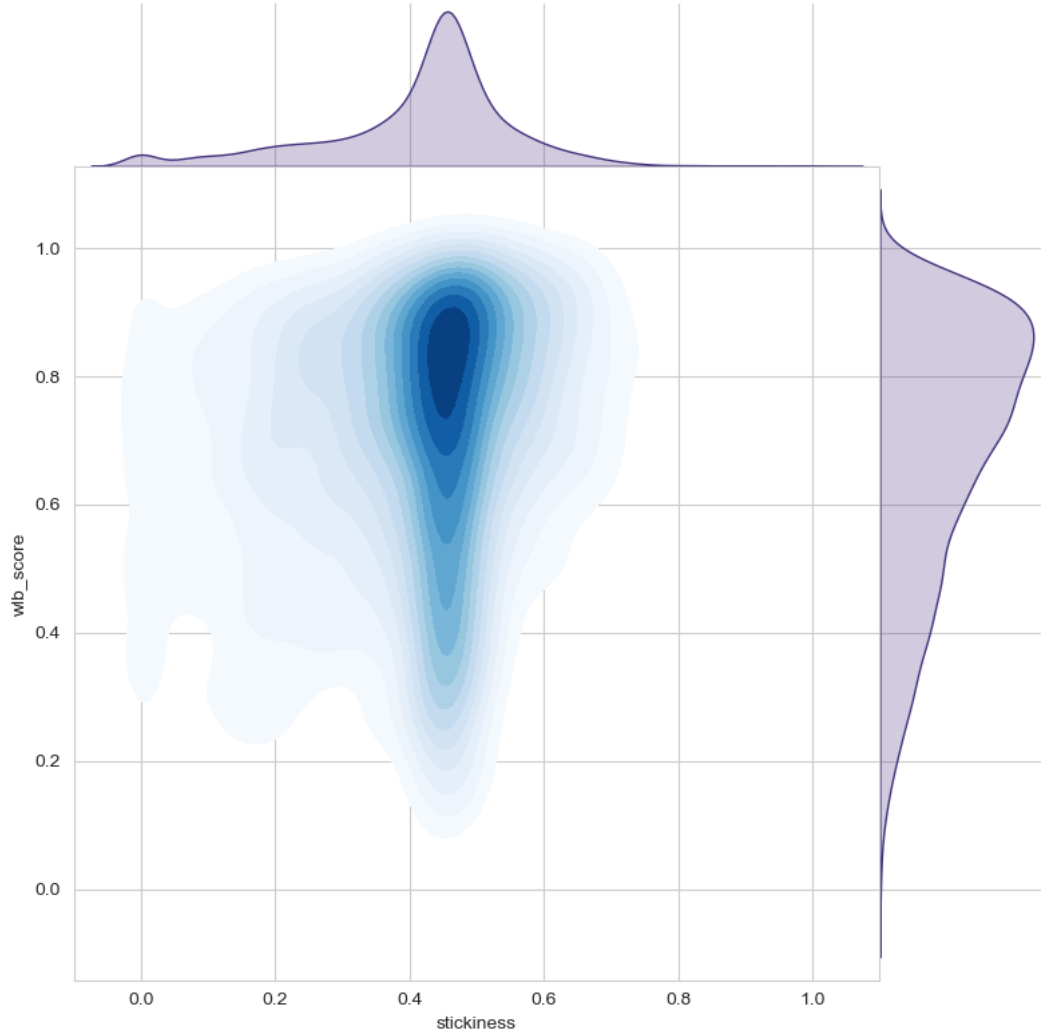
游戏玩时/用户粘性分析



数据长尾分布明显，不利于后续的建模分析；取对数

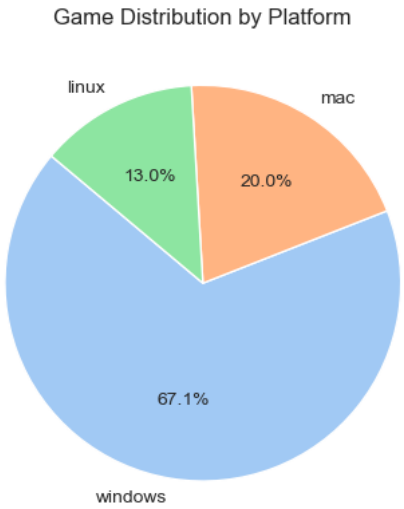


由于指示用户粘性的其他数据难以获取，此处仅以平均玩时和中位数加权平均，因为已经对数变换，二者分布几乎重合，故**等比例加权**得到粘性指标，查看粘性与评分关系

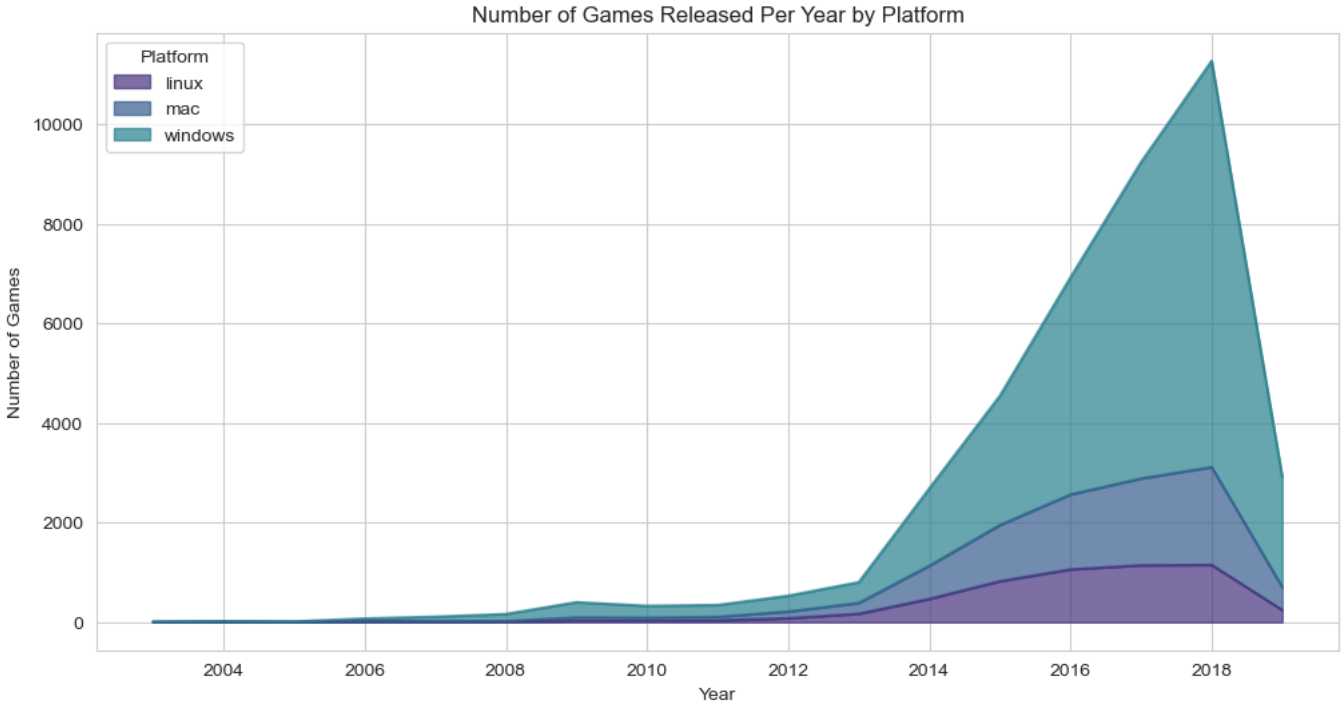


大量游戏因为用户玩时为0，聚集在左侧，但是对于玩时少的游戏，用户评分依然多样化，此外，其他游戏集中在右上角，说明粘性和评分均较高。

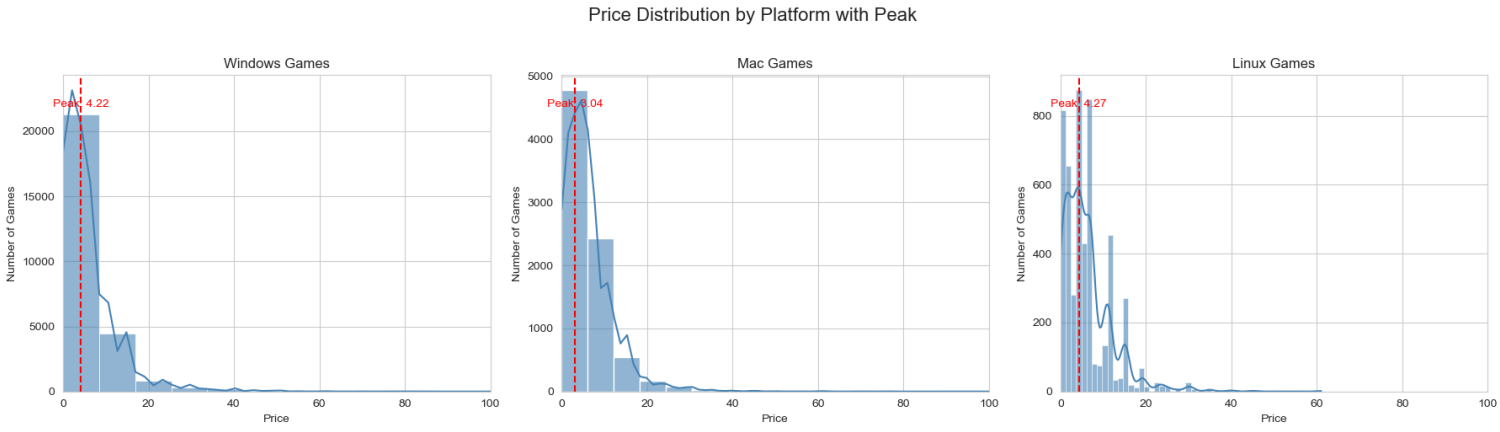
3 平台市场价值评估



平台发行量时间序列

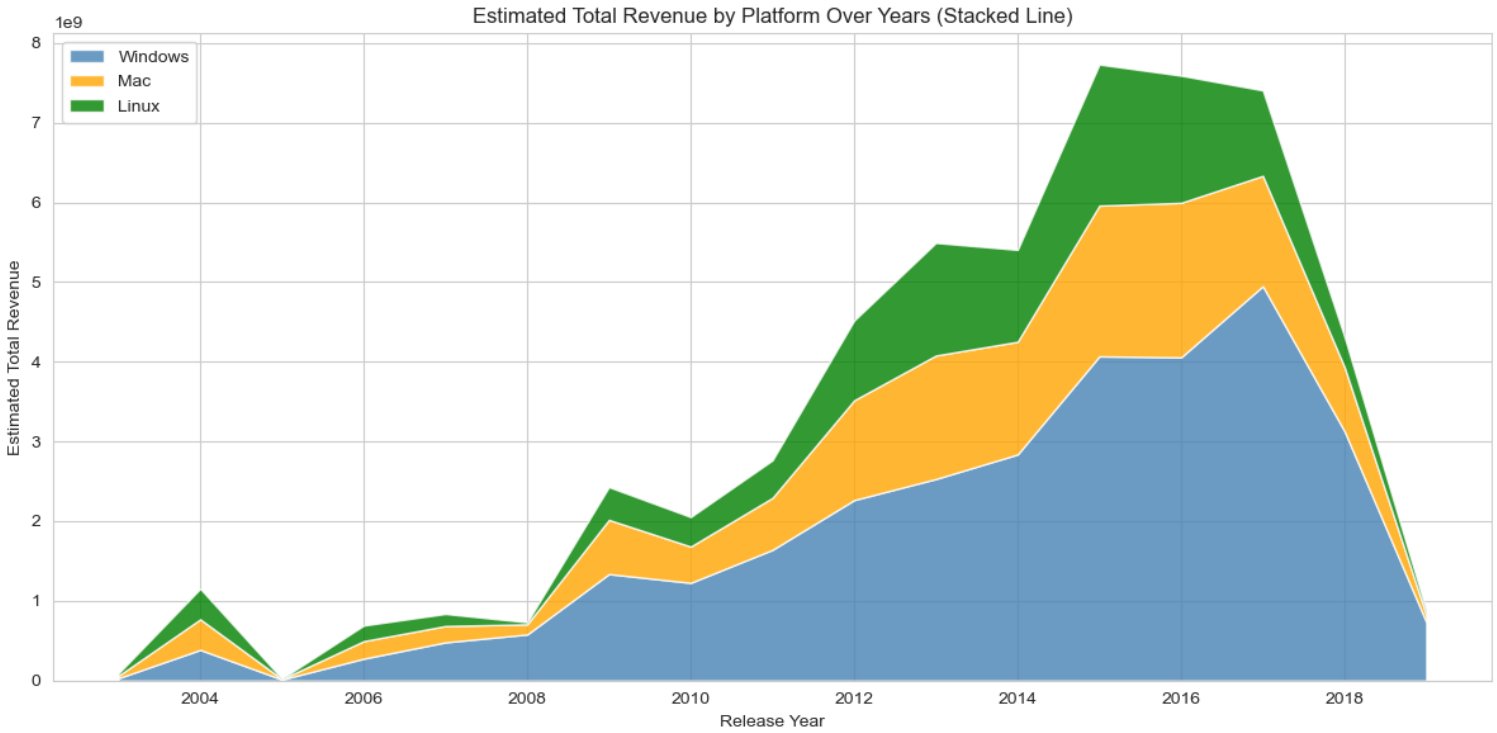


三个平台游戏发行量趋势几乎相同；2013-2018 年游戏发行量激增，2018 年达峰值，后大幅下降。



- Windows 和 macOS 游戏价格分布相似，集中在 0-20，数量随价格先升高后降低，峰值分别在4.22美元和3.04美元。
  - 大部分游戏定价较低，中低价游戏最受欢迎/最多，形成明显峰值
- Linux 游戏价格集中在 0-20，波动显著，峰值在4.27美元。
  - Linux 游戏数量相对较少，价格波动更大，但同样集中在低价区

估计总营收：



三者增减趋势几乎同步！

- Windows 平台：始终是收入主力，占比长期超过 50%。在 2014-2016 年达到收入峰值（约 2.0×10^9以上），即使在下滑期（2016-2018）仍保持相对较高的基数，体现其在游戏市场的核心地位。
- Mac 平台：收入占比仅次于 Windows，在 2012-2016 年进入增长高峰，收入区间约 $0.8\text{-}1.2\times10^9$，与 Windows 形成“双主力”格局；但下滑阶段（2016-2018）降幅明显，收入快速收缩。
- Linux 平台：收入占比最低，整体呈“跟随式增长”。在 2012-2016 年的行业高峰期中，收入从约 0.2×10^9增长至约 1.0×10^9，成为总收入增长的补充力量；但下滑阶段同样受冲击显著，收入几乎归零。

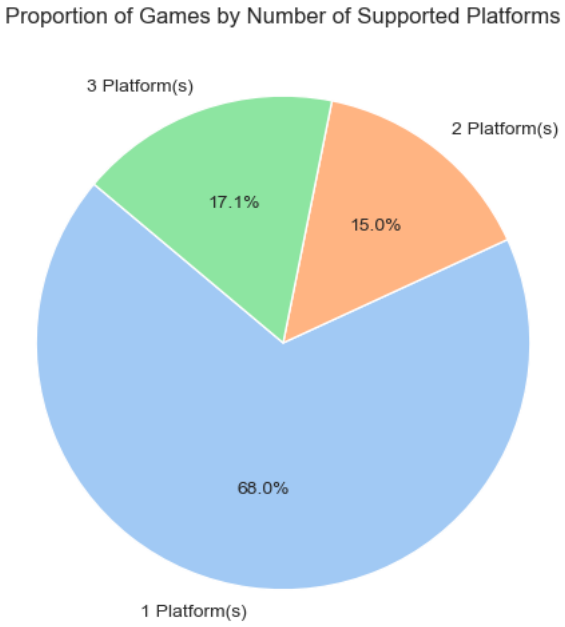
- 2004 年：三大平台均有初始收入，Windows 占比领先，反映其早期在游戏生态的先发优势。

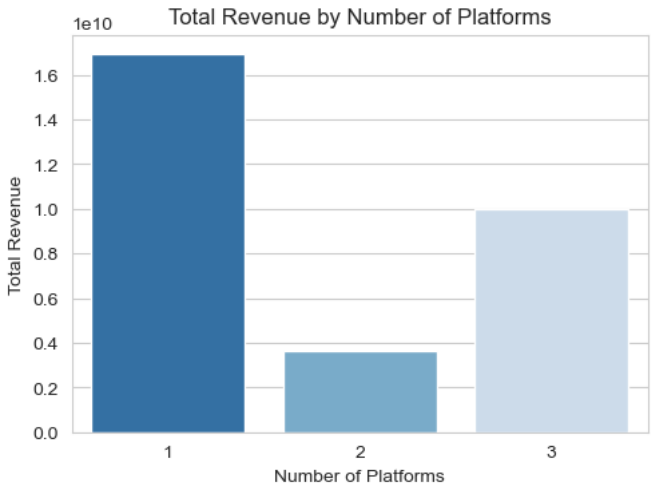
- 2009-2010 年波动：可能与当时游戏行业的产品周期、经济环境或平台技术迭代（如 Windows 系统更新、Mac 硬件升级）有关，导致收入短暂回调。

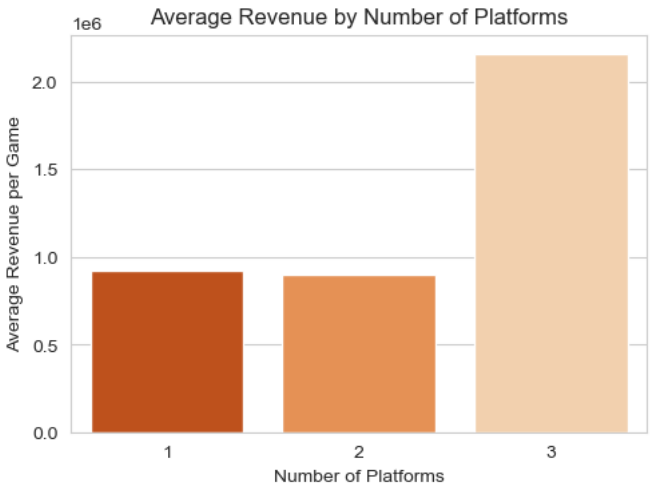
- 2012-2016 年高峰：三大平台同步进入收入黄金期，推测是 PC 游戏品类扩张（如独立游戏爆发、3A 大作频繁上线）、用户付费意愿提升、平台生态完善（如 Steam 对多系统的支持）等因素共同驱动的结果。

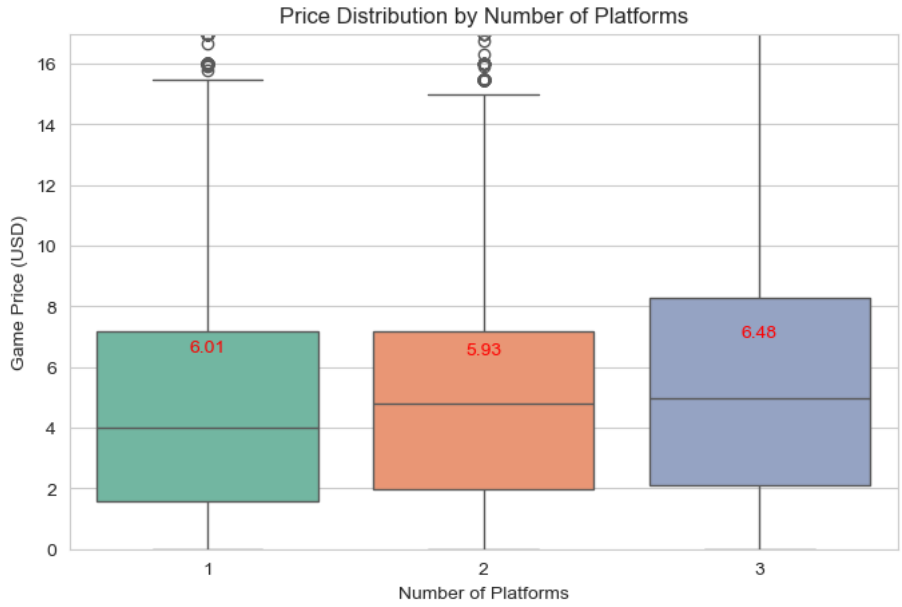
- 2016 年后下滑：总收入的快速回落可能源于市场饱和、移动端游戏分流、PC 游戏创新不足等行业变化，且 Linux 平台的抗风险能力最弱，率先进入收入低谷且无回升趋势。

游戏的平台支持数目





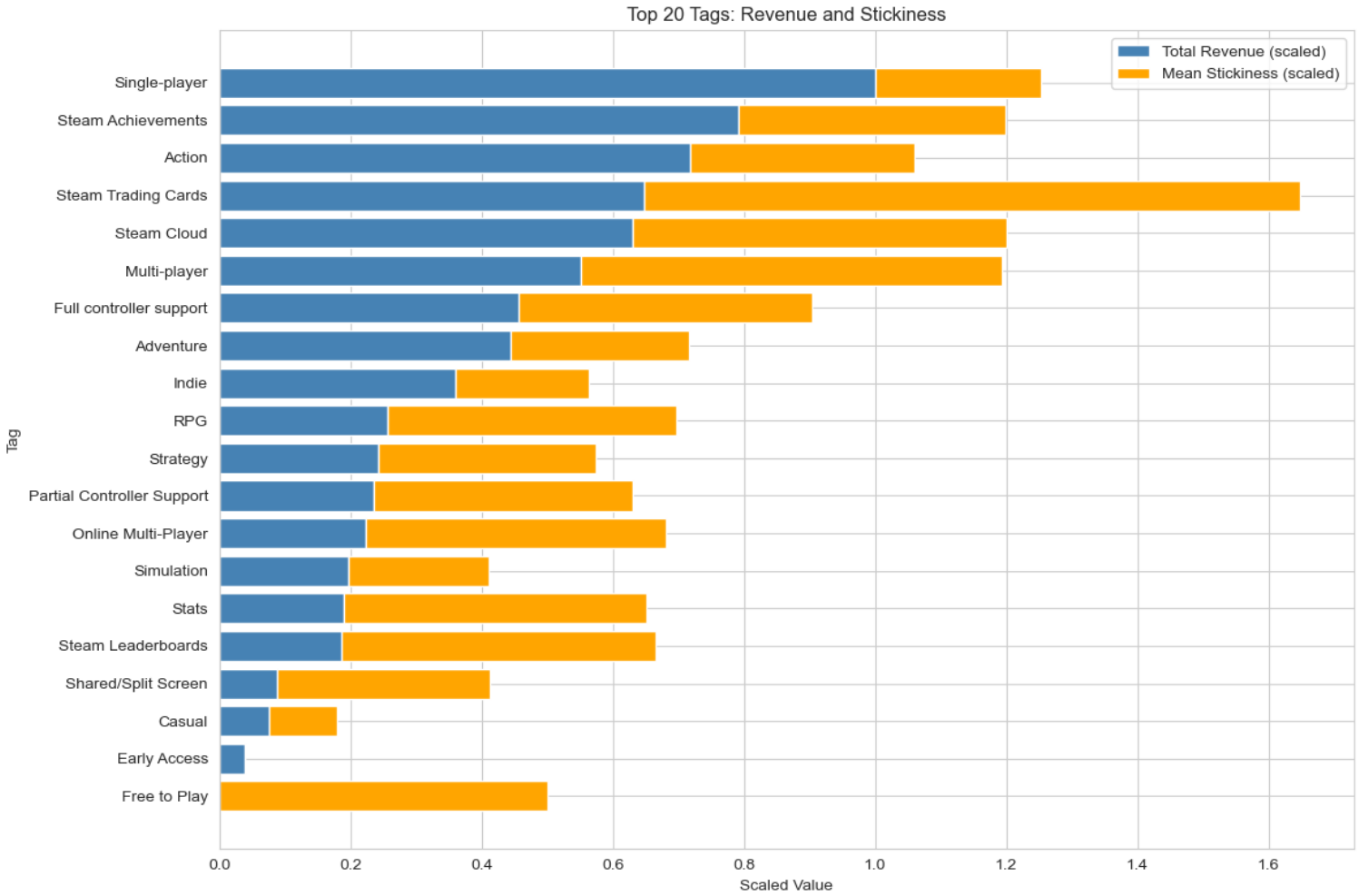




- 发行量与总营收：
  - 单平台游戏占比68%，数量优势带来总营收最高，这是自然结果。
  - 两平台游戏数量少，总营收稍低，但平均收益和单平台几乎相同，说明数量差异是主要原因。
- 三平台游戏：
  - 平均收益明显高出一倍，说明跨三平台能显著提高收益，而不仅仅是增加发行量。
  - 三平台策略对收入提升有实质性效果，支持两平台的效益提升不明显。
- 策略启示：
  - 如果目标是提高营收，直接扩展到三平台比仅增加到两平台更有效。
  - 对于单平台游戏，简单增加一个平台未必带来收益显著提升。

单因素业务关联分析

top20标签的创收和用户粘性差异



- 高营收标签未必对应高粘性（Single-player营收最高但粘性最低），低营收标签也可能有高粘性（如Steam Trading Cards）。游戏的 “赚钱逻辑” 和 “留客逻辑” 是两套体系—— 单人游戏靠优质内容一次性变现，卡牌、多人玩法靠长期互动留存用户
-
  - 单人游戏（Single-player）：聚焦 “内容质量 + 一次性付费”，通过剧情、画面等核心体验驱动高营收，无需依赖用户长期留存；
  - Steam 集换式卡牌（Steam Trading Cards）：靠 “收藏 + 社交” 属性维持高粘性，可作为游戏的 “附加变现与留存工具”，适合在多人游戏、社区型产品中植入；
  - 多人游戏（Multi-player）、动作 / 角色扮演（Action/RPG）：需平衡 “营收（如内购、dlc）” 与 “粘性（玩法循环、社交绑定）”，通过持续更新玩法、搭建社区来维持用户活跃与付费。
- Free to Play（免费游玩）、Indie（独立游戏）等低营收标签中，部分粘性较高（如Free to Play粘性 0.13），说明免费模式、独立游戏可通过 “细分受众 + 差异化体验” 实现用户留存，适合作为 “niche 市场切入点”，后续可通过周边、社群运营挖掘商业价值。

tag的社会网络分析：

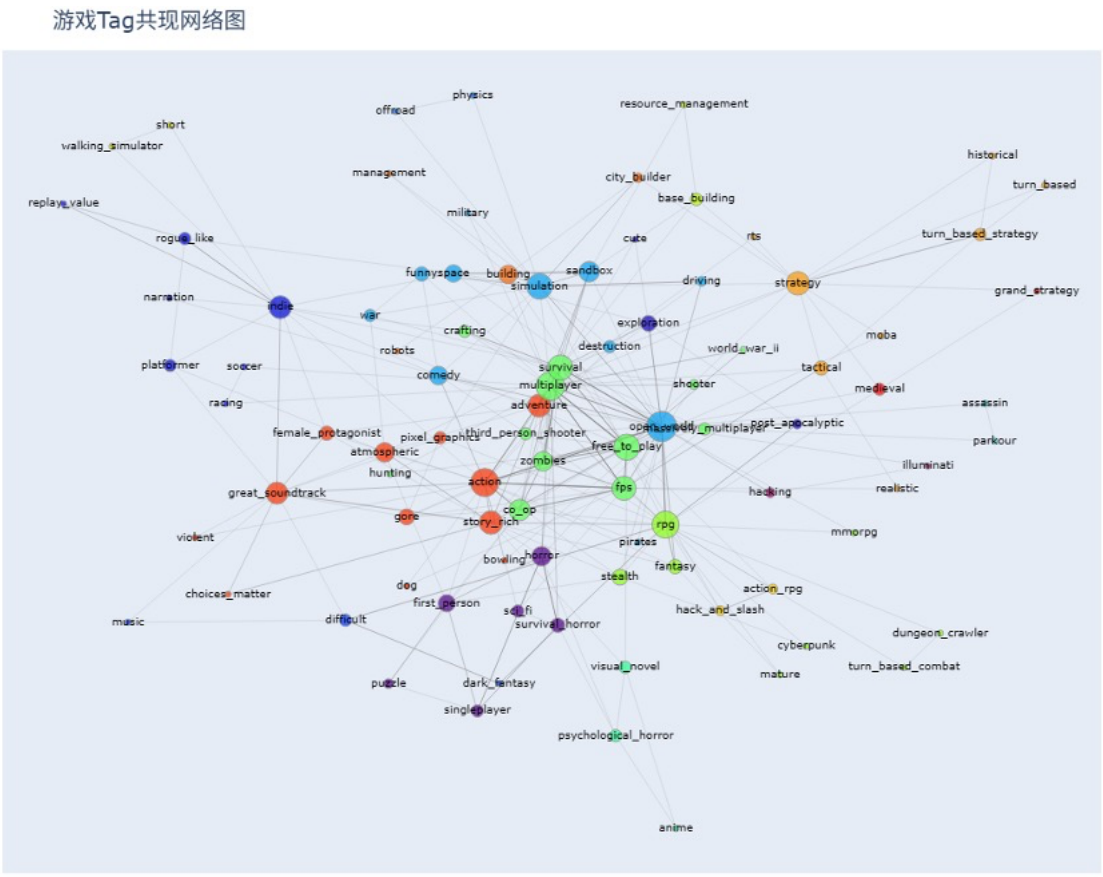
为了使数据更加具有代表性，同时使社会网络的可视化效果更好，需要对数据进行筛选，挑选出更加具有代表性的游戏。这里采用的方法是统计每个游戏的总投票数，选取其中总数大于10000的游戏。由于游戏的总投票数与游戏的玩家数是正相关的，因此这样筛选出来的游戏样本的热度相对更高，更加具有代表性。

在筛选出样本后，还需要对每个样本的tag进行筛选。由于SteamSpy的tag是由玩家参与标引的，无法保证其质量水平，因此，对样本tag的筛选是必要的。这里我们采取的方式是统计每个tag被投票数，选择排名前三的tag来代表该游戏。

将每个tag视为一个节点，可以对tag的共现关系进行连线来构建社会网络。该可视化统计了tag出现次数与两tag的共现次数。tag的出现次数与网络节点的大小成正比，两tag的共现次数与边的粗细成正比。这样的处理使得社会网络的可视化更加直观，重点更加的突出。

除此以外，我还使用了Louvain社群算法来对游戏进行分划，并用不同颜色进行标注，相同颜色的节点属于同一社群。同时使用Spring_layout算法来调整节点布局，使得社群节点分布的远近更能代表其关系的紧密程度。

最后，为了更好展示社会网络的细节，采用plotly来展示社会网络，实现社会网络的可交互。



社会网络图谱节点的大小表示了该tag的出现频率高低，节点越大，表明游戏市场上该类游戏越多。

两节点间连线的粗细表示了两个tag的共现频率的高低，线条越粗，说明两个tag越相互适配。

节点的颜色反映了Louvain算法对社区划分的结果，相同颜色的节点被划为一个社区。

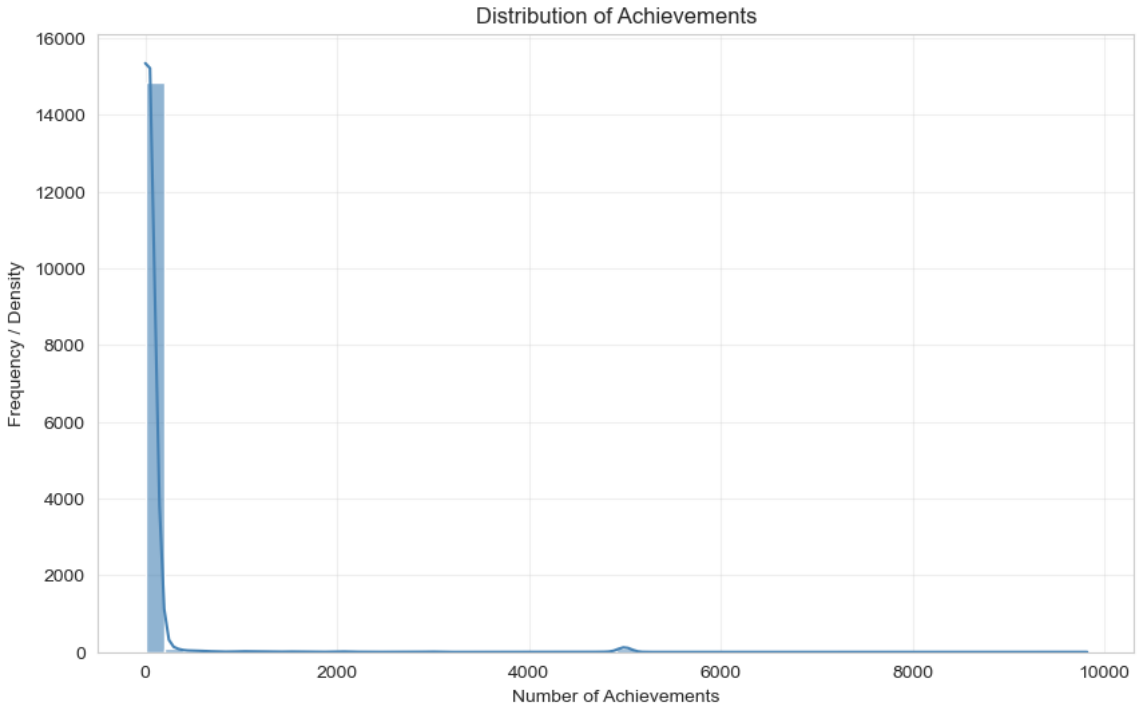
节点可以被分为中心节点与边缘节点，中心节点tag主要描述了游戏的主要玩法、类型等属性（如FPS、开放世界、动作游戏等），边缘节点则反映了游戏的具体主题（如历史、战争、科幻、机器人等）。中心节点与边缘节点、以及分划出的社区关系共同揭示了游戏的玩法与主题的组合情况。（如RPG类与赛博朋克、Dark Fantsy等主题的组合）；

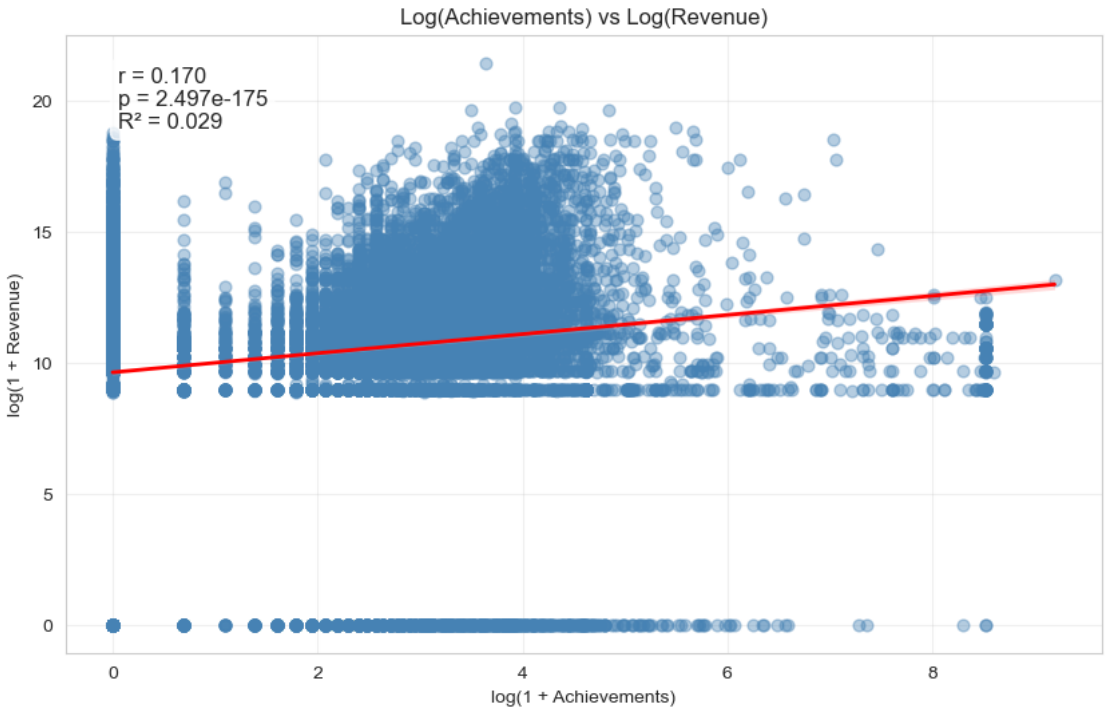
通过社会网络分析可以将游戏大致分为以下几个类群：动作游戏、恐怖游戏、RPG游戏、开放世界沙盒游戏、多人FPS游戏、策略游戏、独立游戏等。同时也存在如跑酷、视觉小说、徒步模拟等更为小众的游戏类型；

社会网络图谱的存在多个社群交错分布的位置，在这些交错的区域中，更有可能出现融合性的游戏，中心位置描述游戏玩法类型的节点交错分布尤为明显，说明当下的游戏在玩法上相对多元化。

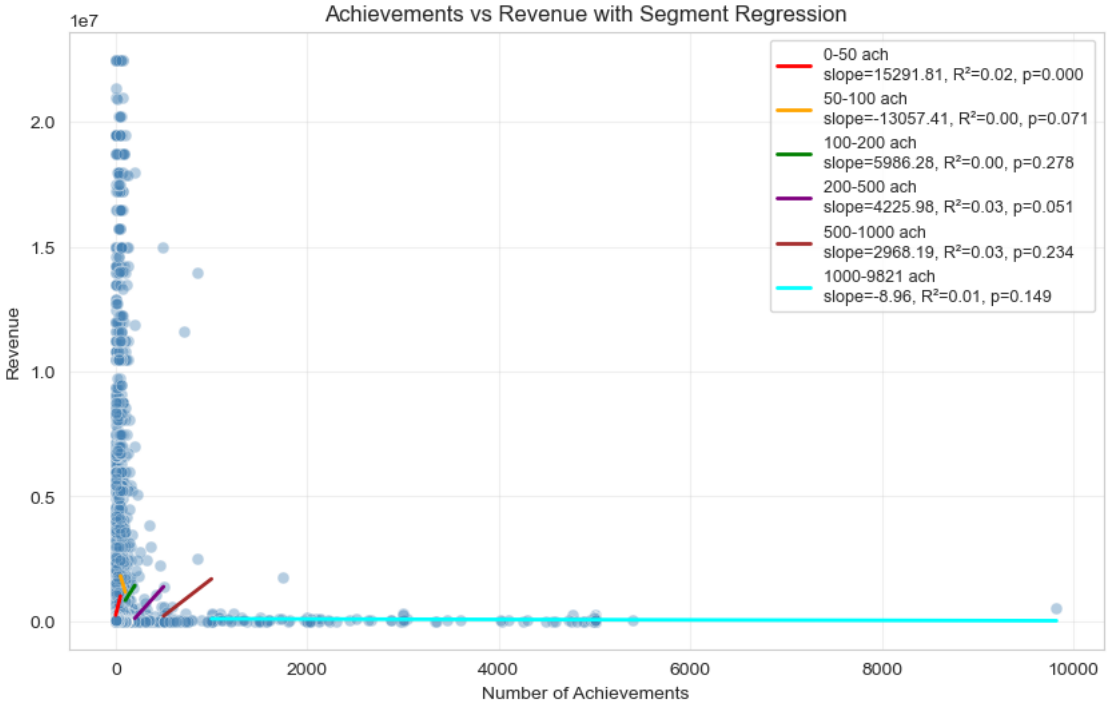
社会网络图谱的边缘位置存在一些相对孤立的游戏类型（如跑酷、视觉小说等），这类游戏在玩法上相对独立，主题上也相对单一，有待未来在其玩法、主题等方面的挖掘。

成就数量和玩家粘性差异





可以看到整体上，二者是存在显著正相关性的（但是经济学上显著性微弱），下面分段回归：



- 仅在 “成就数量极少” 的区间存在微弱正相关，但整体解释力极低；
- 绝大多数区间内，成就数量与营收无明确线性关联，即 “成就数不是营收的核心驱动因素”；
- 过度堆砌成就（1000 个以上）甚至可能对营收产生微弱负面影响（虽然统计学上不显著）
- 简言之，游戏成就系统对营收的作用非常有限，核心仍需依赖游戏本身的内容质量、玩法设计等因素，成就仅能作为辅助性的用户体验优化工具，而非增收利器

游戏热度预测模型特征体系

下面是经过主成分筛选后的特征体系：

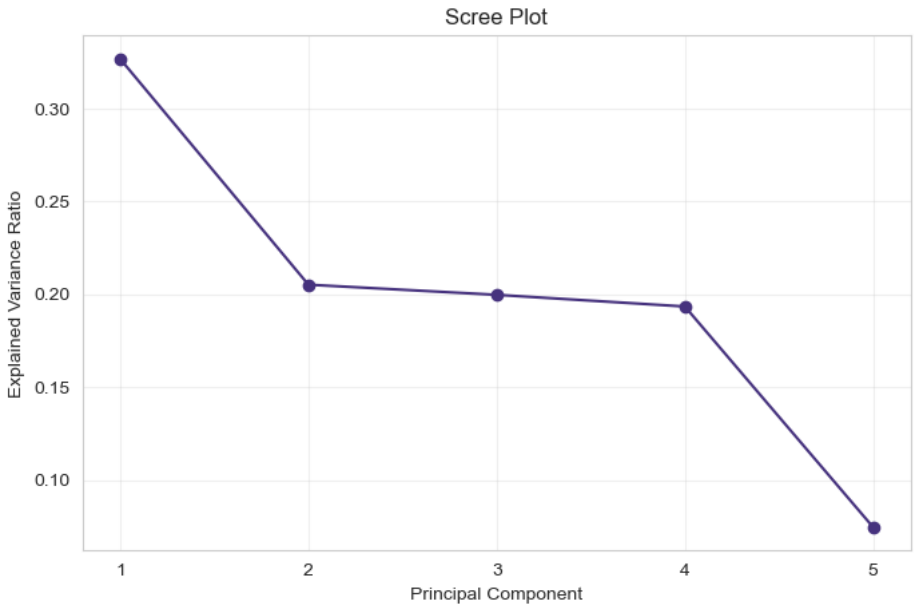
类别	列名	用途 / 作用	处理建议
基本信息	price	游戏定价，影响玩家购买意愿和初期热度	对数变换或标准化
	platforms	支持的平台数量（1/2/3），潜在玩家量	直接计数或独热编码
	owners_est	玩家规模估计（区间中点），反映市场基础	对数变换或标准化
用户行为	stickiness	综合用户粘性指标，反映玩家活跃度	标准化
	achievements	游戏成就数量，间接反映可玩性	对数变换或标准化
	wlb_score	好评率或 Wilson Score，反映用户满意度	标准化或直接使用
游戏属性	tags_aggregated	核心标签，反映玩家兴趣和玩法偏好	选前 N 个热门标签，独热编码或 embedding

衍生特征	early_growth	早期玩家增长速度 = $owners_{est}/daysSinceRelease$	标准化（WLS对量纲敏感）

核心思路

- **行为特征**（stickiness、achievements、wlb_score）直接捕捉玩家真实行为和热度。
- **属性特征**（price、platforms、top_tags、developer）反映市场潜力和玩家偏好。
- 保留最核心特征，降低维度，保证模型可解释性和易操作性。

基本信息



取前三个主成分，即可达到70%左右解释率

在原始数据中，stickiness（用户粘性）存在大量零值，若直接用于建模可能导致以下问题：

1. 对数变换不可行：

由于 log(0) 不存在，零值会引发数值错误或被强制舍弃，造成数据样本损失。

2. 模型不稳定性：

含有大量零值的特征在标准化或回归时会导致方差极低，从而降低模型对该变量的灵敏度。

3. 避免“假零”误导模型：

部分 stickiness = 0 的记录可能并非“完全无粘性”，而是由于数据采样、统计口径等原因产生的“技术性零值”。

因此，为了保证对数变换和建模的数值稳定性，同时不影响整体分布结构，在 stickiness 上加小常数（$1e^{-3}$）：

游戏属性

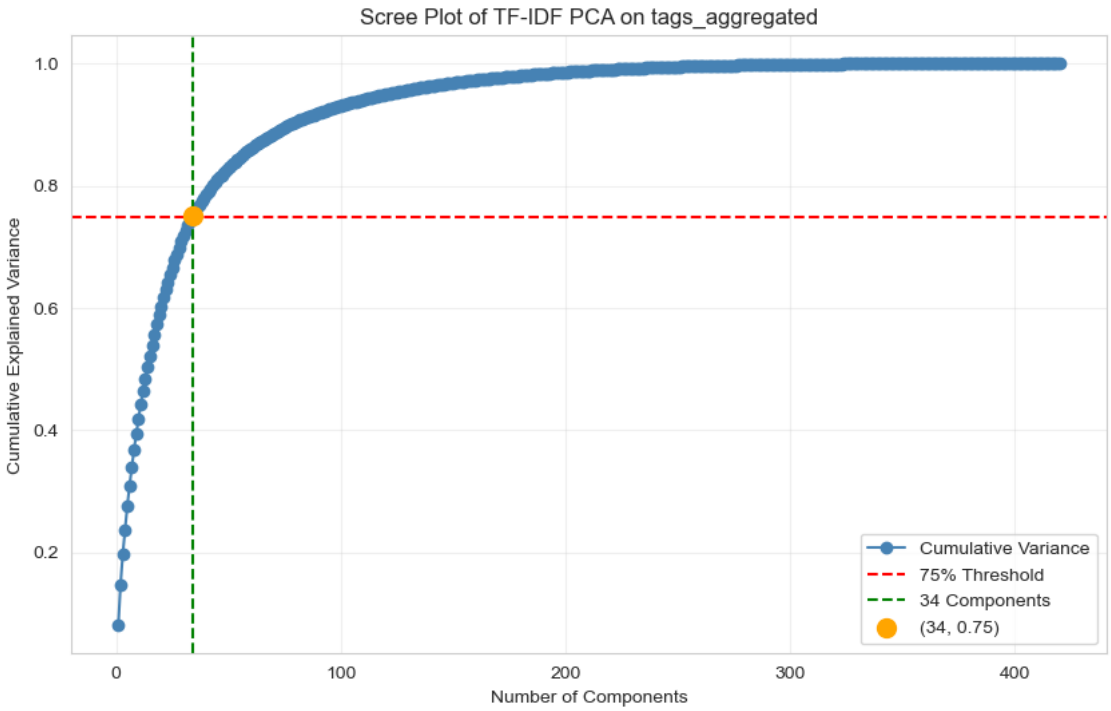
补充硬件配置的要求(来自steam_requirements_data.csv)

	appid	name	release_date	english	developer	publisher	platforms	required_age	categories	genres	...	tag_pc_30	tag_pc
0	30	Day of Defeat	2003-05-01	1	Valve	Valve	windows;mac;linux	0	Multi-player;Valve Anti-Cheat enabled	Action	...	-0.036409	-0.113

1 rows × 78 columns

游戏标签方面，总标签数超过万级，算力不支持。

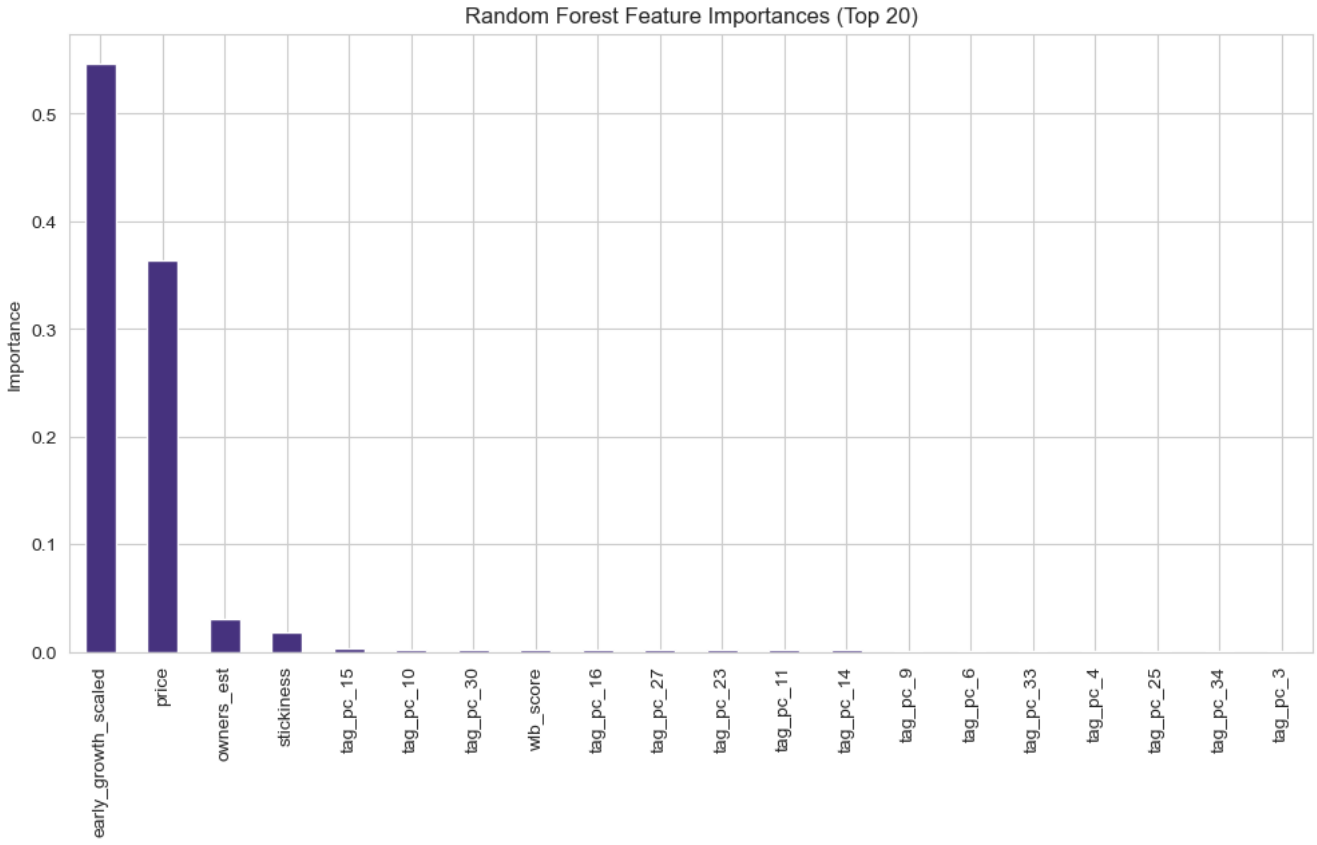
采取TF*IDF法将标签关键词向量化，使用PCA降维。

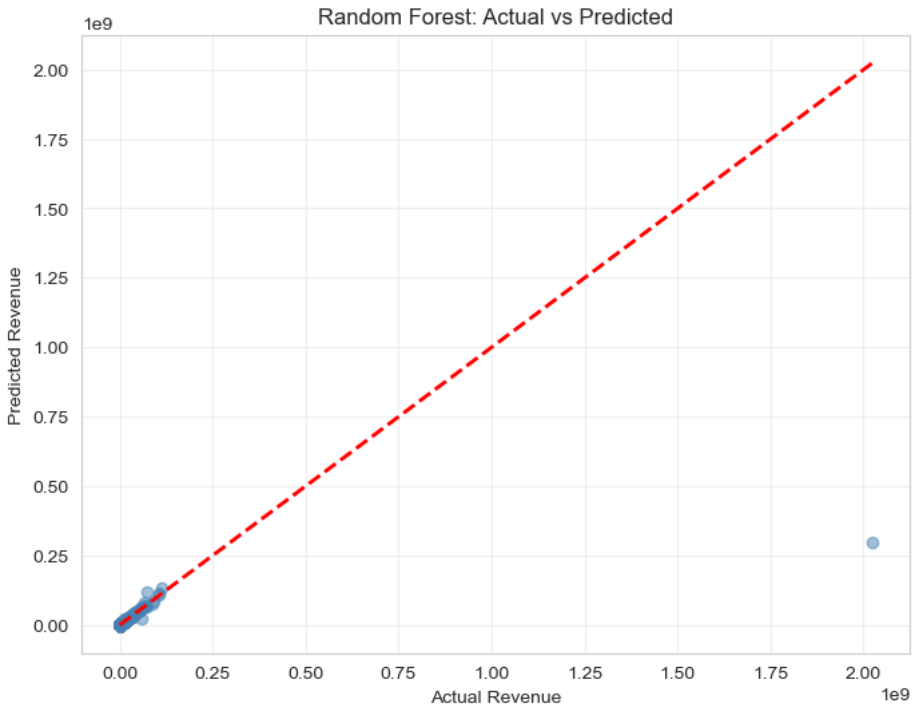


选择保留前34个主成分，方差解释率达到75%

衍生特征：即early_growth标准化后，不赘述

基于以上特征，使用随机森林回归预测





$R^2: 0.296, MAE: 372694.79$

对于随机森林模型而言，此预测效果不容乐观，初步可以怀疑是当前数据特征偏结构化，噪声不低，缺少足够“有信息量”的特征。除此之外，RF 最小化的是 **MSE**，而不关心异方差性。而在下面的分析中可以看见，残差确实是不满足同方差的(异方差性非常显著)

而WLS **显式利用权重建模异方差**，在此场景下有优势。

几乎所有标签的权重都非常小，为了可解释性，选择四个核心特征，线性回归建模(WLS)

WLS模型整体表现

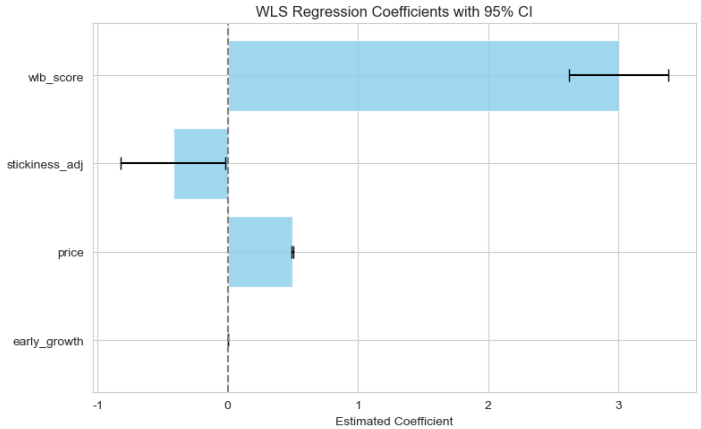
- **R² = 0.515**：模型解释了约 51.5% 的 log(revenue) 变动，四个特征对游戏热度有较强解释力。
- **MAE = 2.89**、**RMSE = 4.13**：对数收入尺度下的平均误差和均方根误差，说明拟合整体稳健。

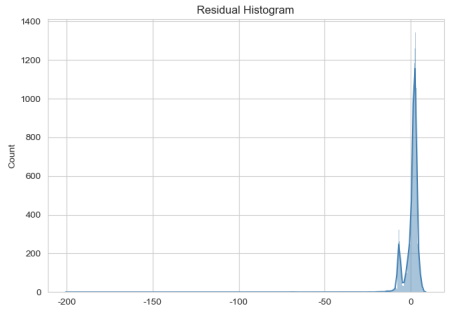
> 模型比随机森林可解释性更强，适合给业务提供具体策略参考。

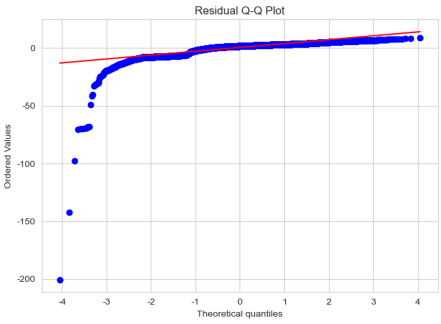
回归系数及业务含义

特征	系数	含义	业务解读
const	5.131	截距	当其他特征为 0 时，log(revenue) 的基准值
early_growth	-0.0002**	对 log(revenue) 的负向影响	系数虽显著，但值很小，早期增长对长期收入影响有限
price	0.498**	对 log(revenue) 的正向影响	游戏价格越高，玩家购买意愿越强，收入明显增加
stickiness_adj	-0.429**	对 log(revenue) 的负向影响	高粘性可能集中在小众或免费游戏，收入未必高
wlb_score	2.997**	对 log(revenue) 的正向影响	好评率越高，玩家满意度越高，直接提升收入，是最强行为驱动因素

> 结论：**价格和好评率是收入的关键驱动因素**，stickiness 与 early_growth 影响较弱。







White Test for Heteroskedasticity

Statistic	Value
LM Statistic	27020.7245
LM p-value	0.0000
F-Statistic	1206364.0771
F p-value	0.0000

业务启示

1. **定价策略**：合理提高价格可直接提升初期收入，但需结合市场和标签偏好优化。
2. **用户评价管理**：提升好评率（wlb_score）最显著，可通过优化内容、社区互动、早期运营干预实现。
3. **粘性与增长**：高粘性和早期增长未必立即带来高收入，但可能影响长期留存和口碑传播，可作为次级指标。
4. **可解释性优势**：相比随机森林，加权线性模型可直接量化每个特征的影响方向与大小，便于产品和开发团队参考决策。

虽然不满足正态性假定，但是由于样本量巨大，可以近似正态处理

四、研究局限

（1）由于使用的数据集只收集了截至2019年的steam游戏数据，导致得出的结论等对于当前的游戏数据的适用程度有待进一步验证；

（2）本研究中主要采用以数据为主的分析方法，忽略了对于游戏选择更重要的因素，如用户的游戏经历、主观意向等（当然受数据集所限，没有这方面数据），结果相对较为片面；若能获取玩家活跃时长数据，可验证'成就数量是否影响留存'这一假设。

（3）由于算力问题，导致一些计算量较大的过程只能尽可能精简。

（4）预测方面，由于特征的数目十分少，且tag权重极小，因此预测效果不可避免地受损，不得已采取退而求其次的加权线性回归最小二乘估计（WLS），在一个不错的拟合优度下提供了较高的的可解释性。

五、未来分析方向

1. 考虑时间动态变化

现有数据为静态快照，无法观察游戏上线后玩家增长曲线。建议接入时间序列数据，构建早期销量预测模型，帮助发行商优化营销节奏。
→ 可提前识别潜力产品，减少资源浪费

2. 加入用户行为细粒度数据

报告仅使用拥有者数量，未区分"下载未玩""玩1小时""长期留存"等行为。若能结合Steam Spy或内部埋点数据，可更精准评估真实用户粘性。
→ 避免将"免费领取"误判为成功

3. 尝试进一步分析DLC与本体的协同效应

许多高收入游戏依赖DLC（如《模拟人生》）。建议补充DLC销售数据，研究"本体定价+DLC策略"对LTV（用户终身价值）的影响。
→ 为长线运营提供定价组合依据

4. 针对标签（Tag）间存在语义重叠进一步分离

如"Indie"与"Adventure"高度共现，可能干扰分类效果。可尝试NLP聚类或降维（如LDA主题模型），提炼更纯净的游戏类型维度。
→ 提升用户画像与推荐系统准确性