统计学视角看电子游戏产业————以Steam平台为例

余柏辰

摘要

近年来，电子游戏产业蓬勃发展，全球各地的游戏玩家对于游戏的管理、体验要求越来越高。许多平台应运而生，其中Valve公司的Steam平台作为规模最大的游戏中介平台，已成为许多电脑游戏玩家不可或缺的电脑软件。

本文基于Kaggle平台上爬取的Steam平台的大量数据和调查问卷得到的精确数据，从个体到行业对影响游戏平台上销量的各因素做了尽可能全面而细致的定性和定量分析。本文通过Python对数据进行清洗，然后通过Echarts生成可视化的描述性统计对数据产生初步的认识，在相关性分析中探索影响电子游戏的重要因素，并使用C#通过统计学习模型进行系统化建模。本文提供了微观与宏观观察电子游戏产业的视角，亦基于数据发现了电子游戏产业两极化的现状。

关键词: 统计学　数据分析　电子游戏产业　统计学习　数据可视化

Abstract

In recent years, with the booming development of the electronic game industry, players over the world have increasingly higher requirements for egame management and experience. Many egame platforms have been created, including Valve's Steam platform, as the largest game intermediary platform, which has become indispensable computer software for many computer game players.

Based on the mega data on Kaggle which crawled data from Steam and accurate data obtained from questionnaires, this paper makes a detailed qualitative and quantitative analysis of various factors affecting sales on the game platform from individuals to industries. The paper cleans the data through Python, then generates a preliminary understanding of the data through visual descriptive statistics by Echarts, explores the important factors affecting video games in correlation analysis, and systematically analyzes the data through statistical learning models via C#. This paper provides micro and macro perspectives to observe the computer game industry, as well as a discovery of the polarization of the video game industry.

Keywords: Statistics　Data Analysis　Electronic Game Industry　Statistical Learning　Data Visualization

一、数据获取

1.数据集

作为全球最大的游戏数字商店，Steam自身提供了内涵丰富的API接口，从中我们可以获取到每个游戏的详细信息。此外，第三方平台SteamSpy提供了额外功能的关于游戏受欢迎程度及销售情况等信息。通过对这两个API内容的爬取，我们可以获得大量的游戏数据。

事实上，由于我们项目的重点在于对数据的处理及分析，这里不妨采用Kaggle平台上Nik Davis于2019年获取的一手数据。这不是一份干净的数据，但也包含了更多的信息，因此我们接下来会对其进行包括补全、去重、降维、归一化等一系列数据清洗工作，使之变得更有效。

2.调查问卷

我们在近150位年轻人中发放了问卷，过滤了无效问卷后一共117份有效问卷。相比于过万的数据集这显得微不足道，但对于某些微观问题其可能得出更具针对性的结果。

二、数据预处理

正如前面提及的那样，在这里我们使用Python编写脚本对下载的数据集进行数据清洗，这是由于其有强大的第三方包和Jupyter Notebook，可以轻松实现实时的观察。事实上，由于数据清洗本身就是一个漫长的过程（有说法称数据分析80%的工作在数据清洗），这里限于篇幅不能展示全貌，故仅以steam_data.csv为例展示主要过程。

首先，通过基本的清洗函数及定义的简单过滤函数，粗略地删除了重复的游戏及数据量不足的行列。

```
def init_process(df):    #将具体的处理过程封装在函数中
    df = df.drop_duplicates()    #删除重复的行
    df = drop_null_cols(df)      #清洗过半数据缺失的列
    df = process_name_type(df)   #清洗id或名称为空值的行
    return df
```

接着，我们逐行观察、处理数据，通过降维使得数据信息密度变大。这里选用is_free列和price_overview列作为展示，这是因为我们可以直观地看到“免费游戏”就是价格为零。因此，我们可以筛选出缺失信息，将其补全，并合成一列。

```
def price_process(df):
    def parse_price(x):
        if x is not np.nan:
            return literal_eval(x)
        else:
            return {'currency': 'GBP', 'initial': -1}
    df['price_overview'] = df['price_overview'].apply(parse_price)  #整理字典
    df['currency'] = df['price_overview'].apply(lambda x: x['currency'])
    df['price'] = df['price_overview'].apply(lambda x: x['initial'])
    df.loc[df['is_free'], 'price'] = 0  #设置免费游戏的价格为0
    df = df.drop(['is_free', 'currency', 'price_overview'], axis=1) #合并列
    return df
```

通过一系列数据清洗的过程，我们将数据提供的信息整合到了一起。当然，数据预处理工作不止于此，如数据归一化对于统计机器学习而言也相当重要，我们将在下面的分析中按照需求进行针对性的进一步处理。

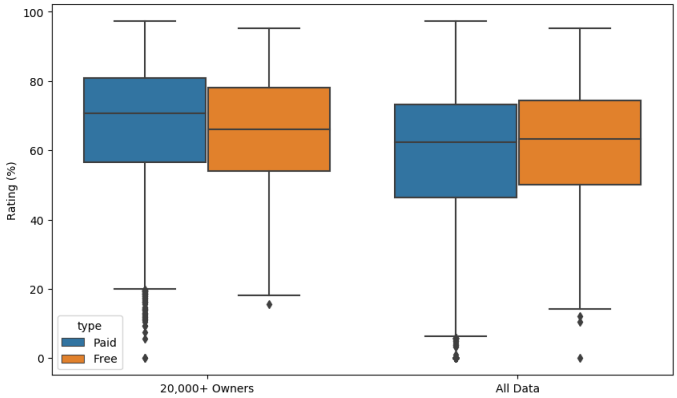
此外，对于收集的问卷数据，由于数据量不大只需简单处理即可，重要的工作是对其进行数据脱敏，以保障数据安全。

三、数据分析与可视化

1.描述性统计

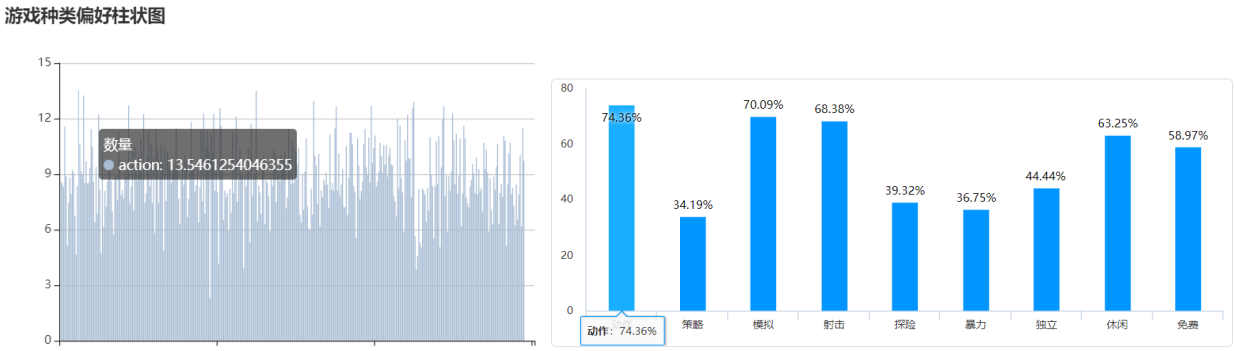
作为最值得关注的数据之一，我们首先生成了游戏拥有量的柱状图。

事实上，由于资本的逐利性，商家可能更关注消费用户的情况。通过下面这个箱线图我们在付费游戏与免费游戏的比较中发现了不同。



与仅考虑20000+拥有量的游戏相比，所有数据中付费游戏的平均评分要低得多。而对于付费游戏和免费游戏哪个更好这样的问题，在游戏人数不同时将得出截然不同的结论。由于免费游戏往往能够吸引更多的用户，同时也会受到更多的争议；而能够拥有两万以上用户的付费游戏一定尤其某方面的过人之处，因此受到欢迎的比例也更大。

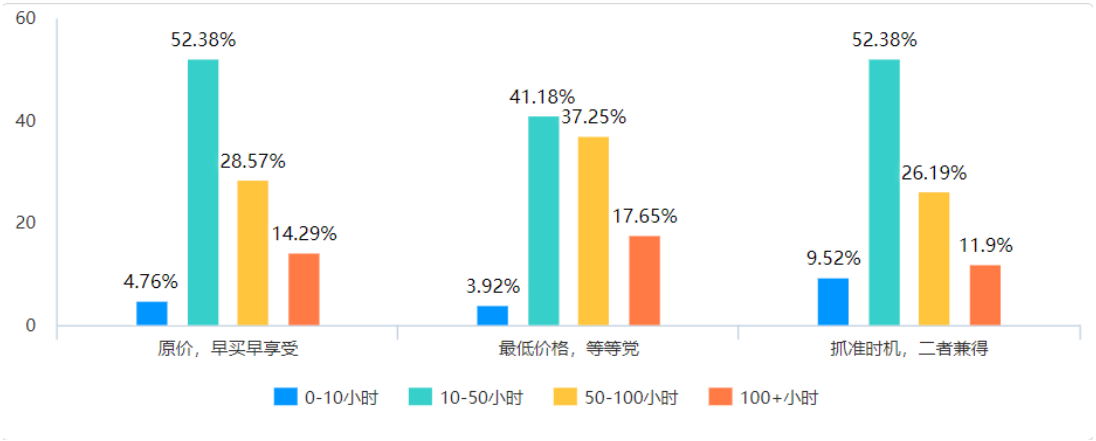
Steam提供了各式各样的游戏类型，而平台数据和问卷数据一致地表明动作类游戏更受用户的喜爱。



对于数据集上爬取的大量数据（左图），我们采用了取对数的方法使效果更为清晰。这种方法在前面游戏拥有量中也曾被尝试，但效果不佳，这是因为在游戏拥有量上数据呈现了明显的两极分化。而对于游戏种类的偏好，虽然在量级方面仍然存在一些差距，但已经是可视的且大多是由于游戏量本身的影响而并非游戏质量的原因。因此，大多游戏类型都是受欢迎的，游戏公司在生产方面则不必追风，完善质量才是王道。

2.相关性分析

商家在决策时一个重要的参考数据即用户的游戏时间。根据问卷的反馈，似乎老玩家更愿意做注重性价比的“等等党”，而冲动消费的玩家往往不会在一个游戏上停留过长的时间。因此我们推测，对一个游戏而言游玩时间过短显然体验效果不佳；而若用户在一个游戏的时间过长则容易导致其不愿接受新游戏。



相应地，我们从平台和用户两个角度来统计分析多少游戏时间更能“塑造”一个优秀的游戏用户。

协方差在统计学中用以衡量两个变量的总体误差，可以通过 $Cov(X,Y)=\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ 计算，其中$\bar{x}$代表X的平均数。如果相关，我们还可以通过 $\rho=\frac{Cov(X,Y)}{S_N(X)S_N(Y)}$ 求得标准化的相关系数（范围在[-1,1]内）。

```
private double calcCov(double[] tmp1,double[] tmp2)
{
    if (tmp1.Length != tmp2.Length)
    {
        Response.Write("<br>计算协方差的两向量维数应一致！<br>");
        return -1;
    }
    double mean1 = calcMean(tmp1);
    double mean2 = calcMean(tmp2);
    double tmp = 0.0;
    for(int i = 0; i < tmp1.Length; i++)
    {
        tmp += (tmp1[i] - mean1) * (tmp2[i] - mean2);
    }
    double Cov = tmp / tmp1.Length;
    return Cov;
}
```

对于商家而言，如上所述，其关注的问题是“游戏时长如何影响购买量”。事实上，分别取游玩时间的平均数和中位数作为游戏时长的向量，计算所得的相关系数分别为0.57和0.16。这和我们前面的猜想不谋而合，即游玩时间多的游戏更受欢迎，但由于边际效应的递减这样的正相关效果并不显著。

而对于用户来说，我们关注的是他们单一游戏时间同总消费账目的关系。二者的协方差为-0.11，意味着几乎没有相关性。这说明游戏产业在游戏时间上表现得不错——有效抑制了游戏成瘾情况的发生。

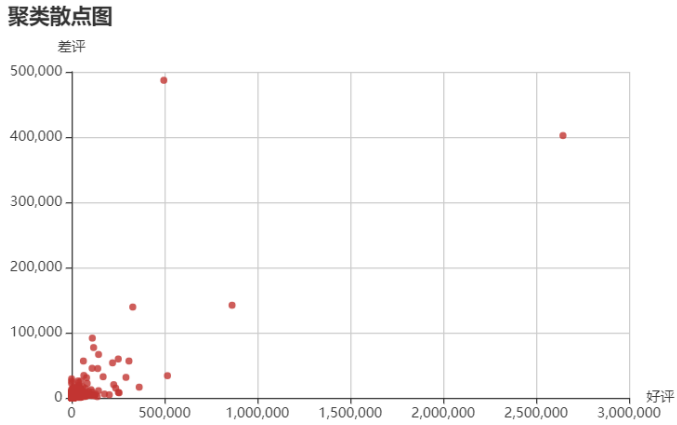
因此，对于游戏生产者而言，也没有必要去刻意地提高用户粘性，生产更多的受欢迎游戏才是上策。

3.统计学习模型

另一个视角来看，其他用户的评价对用户研究往往显得更为直接。因此我们通过用户的评价对游戏进行挖掘，采用K-Means聚类算法构建统计学习模型，依据积极和消极评价两个维度的数据聚为三类，看看哪些游戏备受瞩目，而其中又孰优孰劣。

```
int k = 3;
double[][] rawData = new double[dt.Rows.Count][];
string[][] name = new string[dt.Rows.Count][];
for(int i = 0; i < dt.Rows.Count; i++)
{
    rawData[i] = new double[] { double.Parse(dt.Rows[i][12].ToString()),
double.Parse(dt.Rows[i][13].ToString()) };
    name[i] = new string[] { dt.Rows[i][12].ToString(), dt.Rows[i][13].ToString(),
dt.Rows[i][1].ToString() };
}
int[] cluster = ClassKmeans.Cluster(rawData, k);
ShowClustered(name, cluster, k, 1);
```

==========
Counter-Strike: Global Offensive
PLAYERUNKNOWN'S BATTLEGROUNDS
==========
Team Fortress 2 Dota 2 PAYDAY 2 DayZ Heroes & Generals Rust Grand Theft Auto V No Man's Sky Unturned ARK: Survival Evolved Tom Clancy's Rainbow Six? Siege Fallout 4 Dead by Daylight Z1 Battle Royale Paladins?
==========
Team Fortress Classic Day of Defeat Deathmatch Classic Half-Life: Opposing Force Ricochet Half-Life Counter-Strike: Condition Zero Half-Life: Blue Shift Half-Life 2 Counter-Strike: Source Half-Life: Source Day of Defeat: Source Half-Life 2: Deathmatch Half-Life 2: Lost Coast Half-Life Deathmatch: Source Half-Life 2: Episode One Portal Half-Life 2: Episode Two Left 4 Dead Left 4 Dead 2 Portal 2 Alien Swarm Rag Doll Kung Fu Red Orchestra: Ostfront 41-45 Killing Floor
.....



结合散点图和前面的柱状图我们都能发现，收到最多关注的两款游戏是CSGO和绝地求生。通过Echarts的预览功能我们知道CSGO收获了最多的好评，而绝地求生则是褒贬不一的一款游戏。第二档受到关注的游戏也只有十个左右，绝大多数游戏都只能获得不到10万的评论。

因此，这再次印证了前面提及的游戏产业的两极化。更甚于二八定律，电子游戏领域内部的划分更为清晰，即头部寡头垄断与极大多数的完全竞争产品的结合。事实上，对于大多数生产者和消费者而言，这都是一种能够接受的市场局面。

四、结论

对以Steam为例的电子游戏平台，我们通过个体-商家-市场的三维关系，基于大数据的背景，运用统计学和数据科学的方法得到了诸多结论。

1. 游戏拥有量和关注度分布极不均匀，极少数几款游戏吸引了众多关注，而绝大多数游戏在这方面相差无几。
2. 免费游戏能够吸引更多的用户，同时也会受到更多的争议；而能够拥有较多用户的付费游戏受到欢迎的比例更大。
3. 在不同类型的游戏中，动作类游戏更受喜爱，但游戏种类对游戏本身并不构成决定性因素，质量为王。
4. 对于单个游戏而言，游玩时长与购买量成正相关，但由于边际效应递减等诸多因素并不明显。
5. 对于每个用户而言，每个游戏的游完时间与消费量并无关系。故商家不必可以地提高用户粘性，而应致力于生产更多的优质游戏。

展望

数据分布的极端性，一方面帮助我们看清了产业现状，另一方面也对部分统计计算产生了负面影响。由于统计算法的局限，我们选择了可解释性较强的K-Means算法建模进行机器学习，或许可以构建更丰富的统计学习模型。

此外，对于高维数据也有不同的方式进行降维，进而从不同角度展开数据分析和可视化也是值得期待的。