> 23S_BigData_Project
>
> 爬虫、部分数据清洗及部分数据可视化
>
> By 21301117 周书扬

1. 数据获取

1.1 历史价格数据

网站：`[Steam Price History & Sale History](https://steampricehistory.com/)`

由于 Steam 官方并未提供历史价格数据的获取，因此选则该第三方网站进行爬取，共计爬取**10551**条数据。

1.1.1 一级页面

首先观察目标网页`https://steampricehistory.com/popular`结构为简单的二级结构，每个一级页面由50个二级页面对象组成，二级页面中要的表格即为我们需要爬取的数据。翻页后观察到一级页面的翻页操作可通过更改`https://steampricehistory.com/popular?page=X`后的X值实现，因此首先我们可以利用一个简单的循环遍历每个一级页面。

1.1.2 二级页面

然后我们观察二级页面，发现url中有一串数字id，以游戏`Portal(传送门)`为例，对应的二级界面url为`https://steampricehistory.com/app/400`，其中**400**即为该游戏在steam内部对应的游戏id，而经测试id不是连续的整数，因此从一级界面进入二级界面获取数据仍是较为合理的方式。另外，此id是Steam

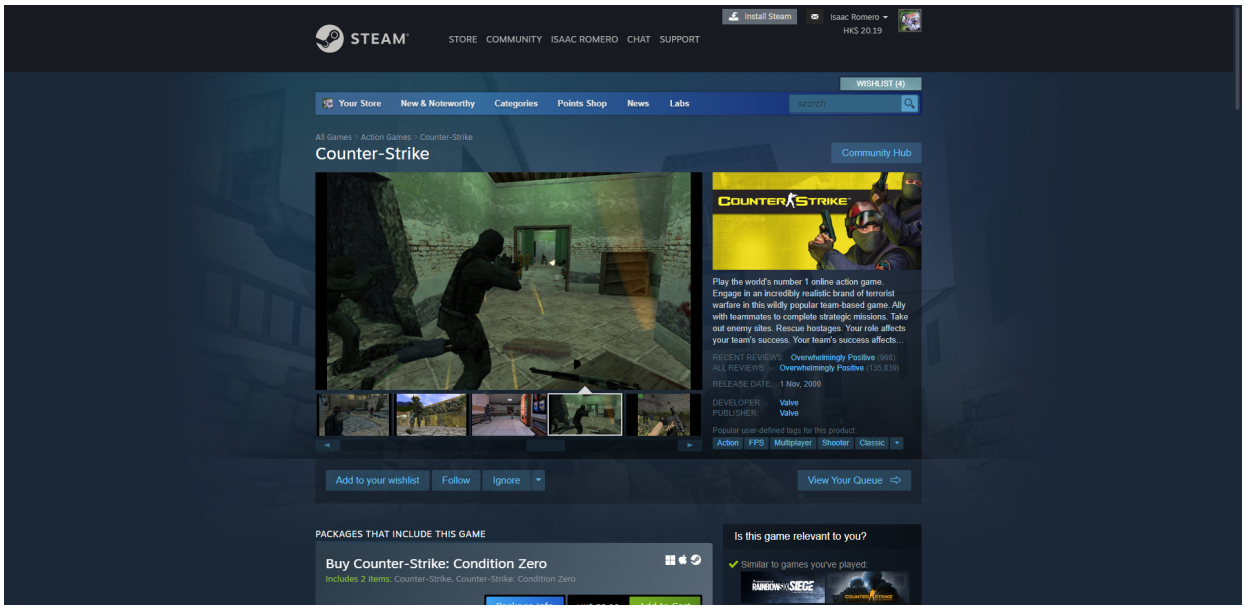
1.1.3 整体实现

使用`request`库发送请求获取网页对象，然后使用`pandas`库读取网页中的表格并存入文件中。另外还使用了`BeautifulSoup`以及一些其他的库实现读取游戏名称和id等操作。同时在爬取前需要添加代理IP和请求头，防止网站禁止访问。

1.2 游戏详细信息

网站：`[Welcome to Steam (steampowered.com)](https://store.steampowered.com/)`

详细信息需要从Steam官网上获取，以**CS1.6**为例：



需要从页面中获取的数据有：

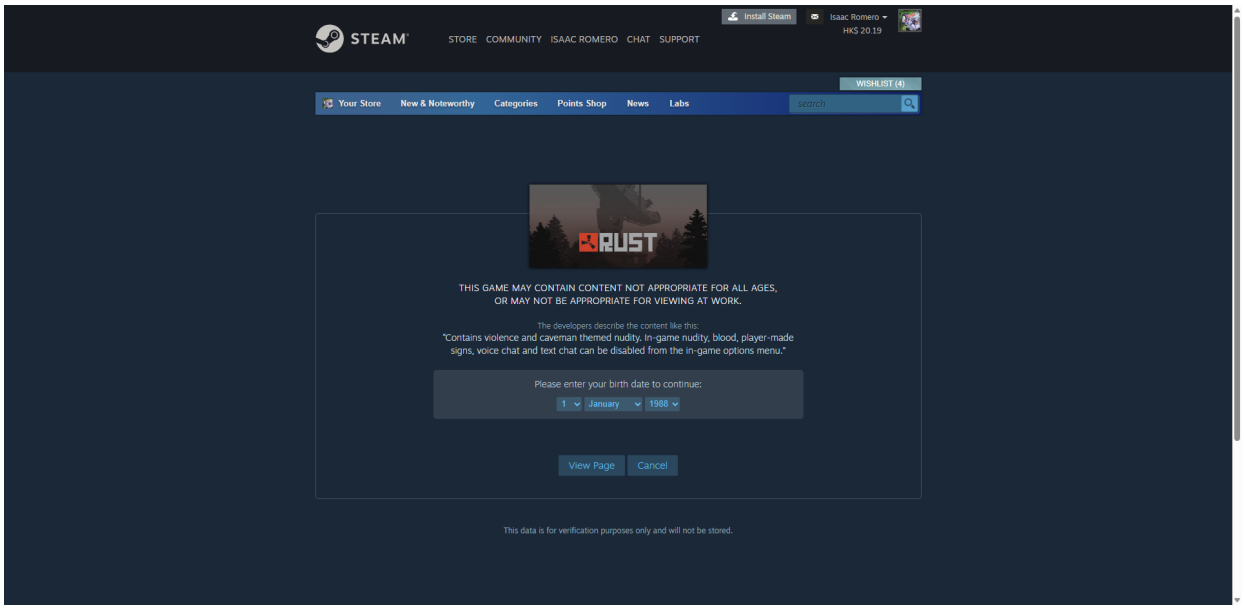
- 游戏名称：Counter-Strike
- 游戏标签：Action FPS Multiplayer
- 游戏描述：Play the world's number 1 online action game. Engage in an incredibly realistic brand of terrorist warfare in this wildly popular team-based game. Ally with teammates to complete strategic missions. Take out enemy sites. Rescue hostages. Your role affects your team's success. Your team's success affects your role.
- 近期评价：Overwhelmingly Positive
- 发行日期：2000/11/1
- 近期好评率：96% of the 978 user reviews in the last 30 days are positive.
- 开发商：Valve
- 评论：一些评论区的评论

1.2.1 爬取思路

利用价格数据中已经保存的游戏id与steam游戏详情页面的url进行拼接，根据url进行爬取，如果遇到锁区或未知情况则跳过爬取

1.2.2 爬取过程中的问题

Steam访问时血腥暴力游戏时会有年龄验证页面或者登录验证界面，因此单纯用`requests`库访问有时会获取不到相应的页面。



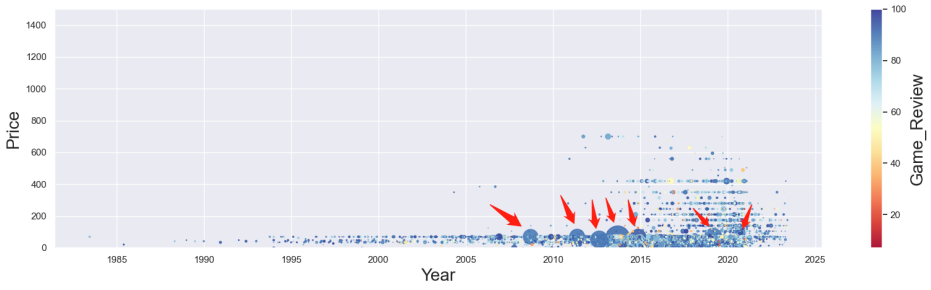
因此需要用`selenium`库结合cookies和模拟鼠标点击获取需要的页面。

2. 数据清洗

1. 删除至少一个字段为空的行
2. 为表添加**价格**属性，具体实现为根据表中游戏id寻找对应历史价格表中的未打折数据添加到记录末尾
3. 对表中列的数据类型进行转换
    1. 将发行日期转换为date64格式
    2. 去除价格列的货币符号并转换为float64类型
    3. 使用正则表达式提取**近期好评率**中的**好评率**和**评价数量**，处理后作为新的两列添加到表格中

3. 数据可视化

使用`matplotlib`库进行可视化。



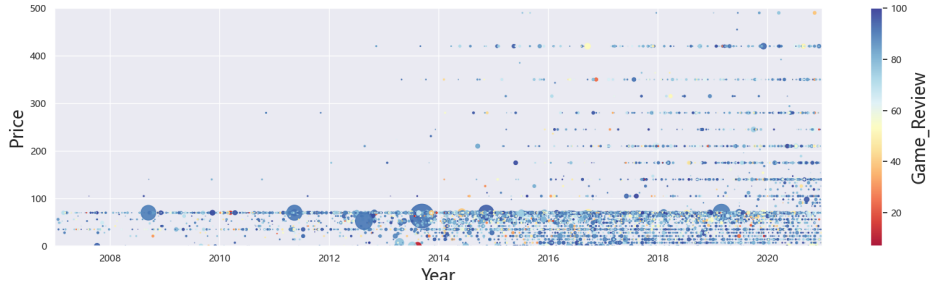
图中横坐标代表年份、纵坐标代表价格（RMB）、散点的大小代表着游戏的评价数量（一定程度上可以反应游戏热门程度）、颜色代表好评率，越接近蓝色好评率越高，越接近红色好评率越低。

可以发现图中大部分游戏集中在2012年之后，可以推测Steam从2012年起开始慢慢被更多游戏玩家和厂商接受，几个比较大的点也基本都集中在2012年后。找出这几个点对应的游戏后也可以看出的确是比较热门的游戏。

B	C	D	E	F	G	H	I	J	K	L	M	N
Link	ID	名字	标签	描述	近期评价	发行日期	期数量好评	开发商	评论	价格	评价数量	好评率
https://s	20900	The Witch	RPGFantas	Become Th	Very Pos	########	89% of th	CD PROJEK	,关于这款	9.99	59501	89
https://s	20920	The Witch	RPGFantas	A time of	Very Pos	########	90% of th	CD PROJEK	求用巫师3	9.99	62129	90
https://s	730	Counter-S	FPSShoote	Counter-S	Very Pos	########	90% of th	Valve	天天被刀,	7.49	78484	90
https://s	107410	Arma 3	ActionMi	Experienc	Very Pos	########	91% of th	Bohemia I	给所有想3	8.99	154926	91
https://s	247730	Nether: I	Survival	Nether is	Mixed	########	40% of th	Nether Pr	这游戏无	9.99	17197	40
https://s	244850	Space Eng	SpaceSand	Space Eng	Very Pos	########	89% of th	Keen Soft	Space, Eng	9.99	78344	89

另外我们爬取的游戏信息中大多数游戏售价都集中在100元RMB以下，整体价格比较亲民。

下面将图标局部放大一下，截取2007年至2021年，价格在0-500元RMB之间的部分



可以看到在80元左右有一条非常明显的分界线，分界线以上比较稀疏，分界线以下非常密集，并且上文提到的几个热门游戏都在分界线上，这些制作精良的作品开发成本同样也较高，因此这些热门游戏的价格也往往代表了行业里较高的水准，代表了玩家对游戏定价的最高接受能力。所以在这条分界线以上的作品数量明显较少，而热门程度也相对较低。

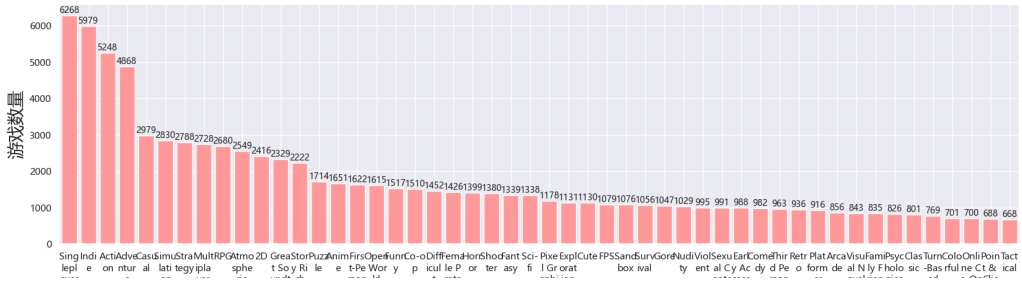
可以看出游戏的平均好评在2012年前波动较大，并且呈现下降趋势，在2012年后开始较为稳定。而平均价格则在2015年开始有了逐年上涨的趋势。

那么为什么会出现这种趋势呢?

- Steam在2009年尝试了打折业务，2010年正式上线了打折业务，让玩家能以更低廉的价格购买游戏
- 2011年，Steam上线了创意工坊，一个为玩家提供的针对游戏的自定义社区，为社区注入了大量活力
- 2016年，Steam上线微信支付宝功能，正式走进国内玩家的视野

可以看出基本每个时间节点都能找到对应的事件解释数据波动的原因

最后，我们再根据游戏的标签统计出出现频率最高的50个标签，统计结果只有**52**个游戏不能被这前50个标签覆盖，因此这50个标签基本可以反应整体的情况。将这50个标签的出现频率绘制成柱状图，得到以下结果:



可以看到前4个标签出现的频率非常高，几乎一多半的游戏都含有这些标签。同时第50名频率就已经下降到了600，而标签总共有441种，因此标签的频率服从一个非常标准的重尾分布。

这里列出排名前10的标签：

```
Singleplayer
Indie
Action
Adventure
Casual
Simulation
Strategy
Multiplayer
RPG
Atmospheric
```