数据科学家的第一个 Project

项目要求：

内容完整程度
可用性（可操作、易操作、美观）
时间先后
先进性

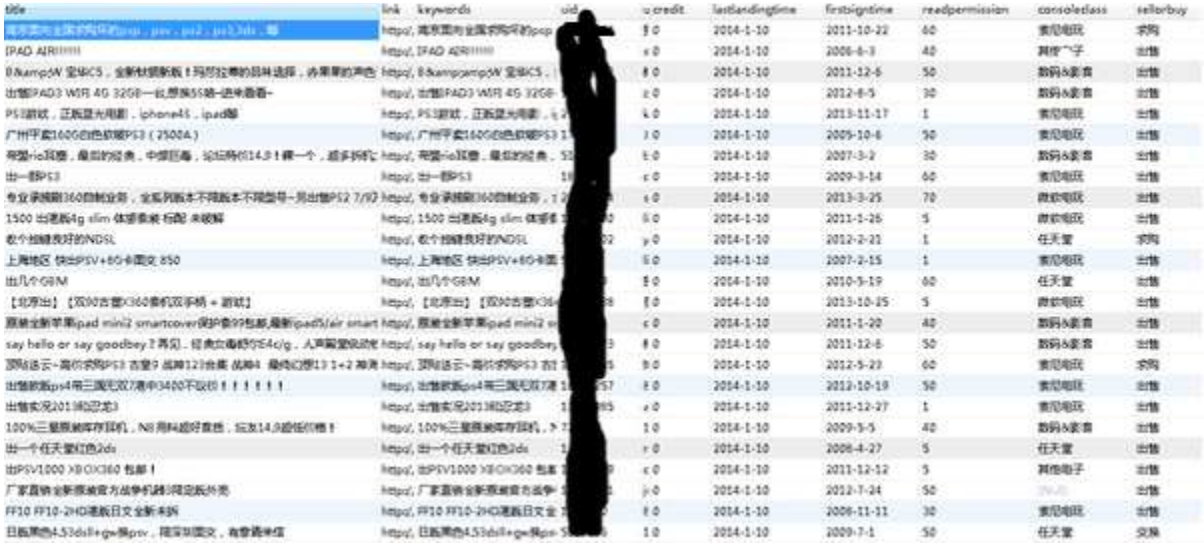
下面给出一个 demo，供实现路径参考

确定项目目标

例如：分析某二手主机游戏交易论坛上的帖子，从中得出其用户行为的描述，为用户进行分类，输出洞察报告。

从互联网获取数据

例如：可以用 python 写个定向爬虫，抓了某个著名游戏论坛的二手区所有的发帖信息，包括帖子内容、发帖人信息等，样子可能如下：



数据清洗

这个模型中的数据清洗，主要是洗掉帖子中的无效信息，包括以下两类：

1、论坛由于其特殊性，很多人成交后会把帖子改成《已出》等标题，这一类数据需要删除：

2、有一部分人用直接贴图的方式放求购信息，这部分体现为只抓到图片链接，需要删除。

后边分析时会再进行一轮清洗。

数据整理

需要把数据整理成可以进一步分析的格式。

首先，给每条帖子打标签，标签分为三类：行为类型（买 OR 卖 OR 换），目标厂商（微软 OR 索尼 OR 任天堂），目标对象（主机 OR 游戏软件）。打标签模式是”符合关键词—打相应标签“的方法，关键词表样例如下：

consolename	consoleactualname	consoleproduct	consoleclass
XBOX360	XBOX360	微软电玩	主机
360	XBOX360	微软电玩	主机
X360	XBOX360	微软电玩	主机
XBOX	XBOX	微软电玩	主机
WII	WII	任天堂	主机
NDS	NDS	任天堂	掌机
小N	NDS	任天堂	掌机
小三	3DS	任天堂	掌机
PS3	PLAYSTATION3	索尼电玩	主机
PS4	PLAYSTATION4	索尼电玩	主机
3DS	3DS	任天堂	主机
三公主	PLAYSTATION3	索尼电玩	主机
PSP	PLAYSTATIONPORTABLE	索尼电玩	掌机
小公主	PLAYSTATIONPORTABLE	索尼电玩	掌机
PSV	PLAYSTATIONVITA	索尼电玩	掌机

从发帖用户作为视角，输出一份用户的统计表格，里边包含每个用户的发帖数、求购次数、出售次数、交换次数、每一类主机/游戏的行为次数等等，作为后续搭建用户分析模型之用。表格大概长这个样子：

uid	postcount	postsale	postbuy	postchange	signrange	lastlandingrange	lastpostrange	userprovince
10002118	1	1	0	0	37	1	5	上海市
10004278	1	1	0	0	38	1	1	不明
10006014	1	1	0	0	38	1	1	上海市
10006251	10	9	1	0	38	1	1	广东省
10008226	1	0	1	0	38	2	3	上海市
10009763	47	47	0	0	38	1	1	福建省
10012602	2	0	2	0	38	1	1	黑龙江省
10012608	1	1	0	0	27	1	1	上海市
1001361	2	2	0	0	65	2	2	黑龙江省
10014302	1	1	0	0	38	1	2	重庆市
10015772	5	3	2	0	36	3	4	不明
10018860	1	1	0	0	38	2	2	安徽省
10023515	1	1	0	0	38	1	1	广东省
10024146	4	4	0	0	30	1	1	浙江省
1002470	2	2	0	0	80	1	2	天津市
10024965	7	7	0	0	38	1	2	广西壮族自治区
10025585	1	1	0	0	38	1	5	江苏省
1002593	1	1	0	0	65	4	4	北京市
10026115	2	0	2	0	38	2	2	安徽省
10027220	10	10	0	0	38	1	1	河北省
10027488	1	1	0	0	38	2	4	不明
10028087	1	1	0	0	38	1	4	广东省
1003	1	1	0	0	60	1	2	北京市

之后这个表的列数会越来越多，因为数据重构的工作都在此表中进行。

整理之后，我们准备进行描述统计。

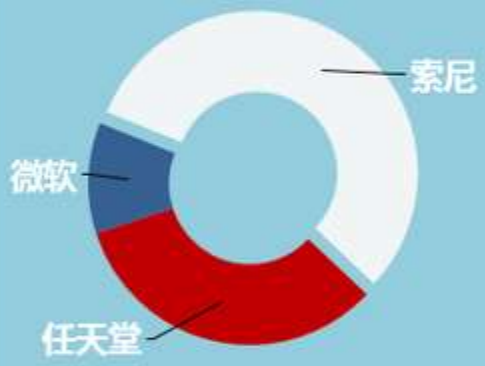
描述统计 & 洞察结论

描述统计在这个项目中的意义在于，描述这一社区的二手游戏及主机市场的基本情况，为后续用户模型的建立提供基础信息。

比如成品图，分别是从各主机市场份额、用户相互转化情况、地域分布情况进行的洞察。

三“国”争锋 一家独大

交易热度对比



每百名游戏主机用户拥有：

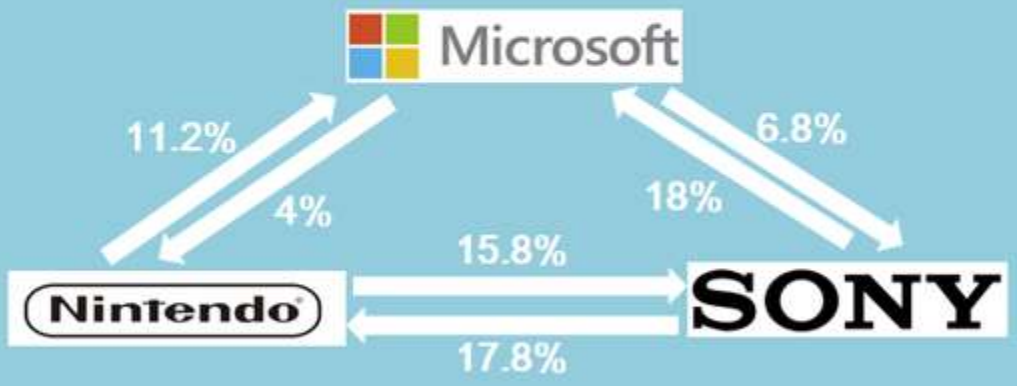
索尼主机60台

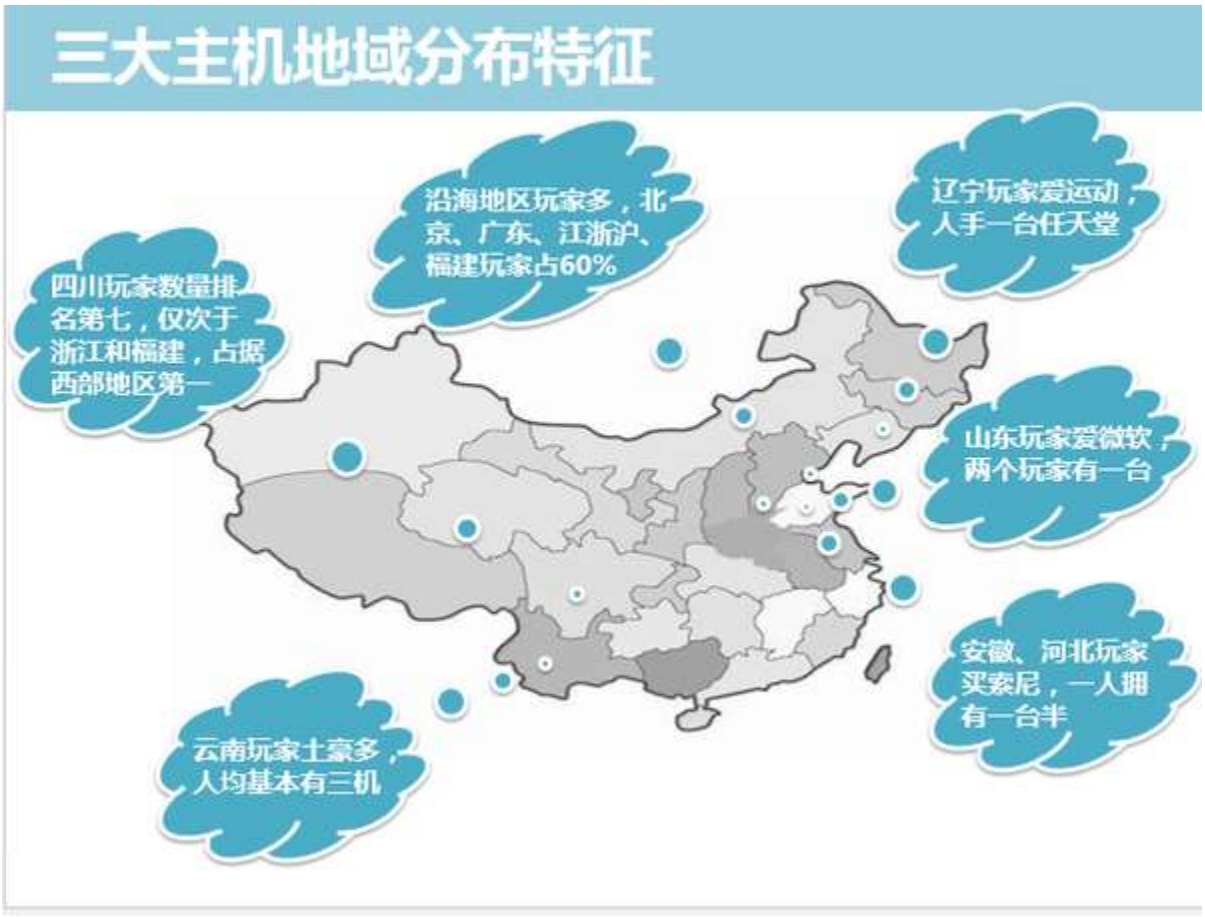
微软主机15台

任天堂主机40台

三大主机用户的流动

各主机的求购者中，均有一部分由其他主机的前拥有者（卖掉原有主机的人）转化而来





选择变量 & 选择算法

因为要研究的是这些用户与二手交易相关的行为，因此初步选择变量为发帖数量、微软主机拥有台数、索尼主机拥有台数、任天堂主机拥有台数。

我们的目标是将用户分群，因此选择聚类，方法选择最简单的 K-means 算法。

设定参数 & 加载算法

K-means 算法除了输入变量以外，还需要设定聚类数，我们先拍脑袋聚个五类吧！

看看结果：

每个聚类中的案例数		
聚类	1	11135.000
	2	18.000
	3	331.000
	4	4.000
	5	1.000
有效		11489.000
缺失		.000

选择变量 & 选择算法 &设定参数 & 加载算法 &重构变量

用原始值来聚类的结果不太好，那么我把原始值重构成若干档次，比如发帖 1-10 的转换为 1,10-50 的转换为 2，依次类推，再聚一次看看结果。

每个聚类中的案例数		
聚类	1	7700.000
	2	1801.000
	3	1169.000
	4	503.000
	5	316.000
有效		11489.000
缺失		.000

聚成四类试试：

每个聚类中的案例数		
聚类	1	5895.000
	2	839.000
	3	1294.000
	4	3461.000
有效		11489.000
缺失		.000

结果测试

测试过程中，很重要的一步是要看模型的可解释性，如果可解释性较差，那么打回重做……

接下来，我们看看每一类的统计数据：

聚类类别	1	2	3	4
平均发帖数	1.71	9.85	1.77	1.84
平均微软主机拥有量	0.04	0.78	1.54	0.01
平均索尼主机拥有量	1.35	6.44	0.08	0.07
平均任天堂主机拥有量	0.10	2.28	0.10	1.71
该类别用户数	5895.000	839.000	1294.000	3461.000

这个表出来以后，基本上可以对我们聚类结果中的每一类人群进行解读了。结果测试通过！

输出规则 & 模型加载 & 报告撰写

这个模型不用回朔到系统中，因为仅仅是一个我们用来研究的模型而已。因此，输出规则和模型加载两步可以跳过，直接进入报告撰写。

聚类模型的结果可归结为下图：

最后附送几张各类用户发帖内容中的关键词词云图：


索尼Fan用户关键词
PSV
PS
ps
游戏
psv
出售
PSP
白色
黑色
面交
中文
北京
破解
上海
微软Fan用户关键词
360
xbox
XBOX
Xbox
PS
求购
一台
游戏
出售
主机
双65
slim
SLIM
白色
黑色
无线
破解
任天堂Fan用户关键词
DS
DSLL
出售
ds
游戏
上海
求购
面交
猎人
白色
一台
系统
黑色
dsll
怪物
wii
MH
口袋
二手商人用户关键词
PS
PSV
PSP
psv
ps
XBOX
360
游戏
出售
求购
黑色
白色
破解
中文
北京
上海

做这份报告用到的工具：


分析工具
• Python——数据抓取
• MySQL——数据整理统计
• Excel——图表制作
• SPSS,R——数据建模可视化