Q2 实验报告

1854116 朱明志

目录

1 数据集描述

给定一份机顶盒数据集，其中，一个机顶盒卡号代表一户家庭，并将一户家庭作为一个用户。该数据集包括用户收藏记录和用户常看直播频道记录样例。两种数据集的具体格式如下：

1） 用户收藏记录

```
{
"CODE"："媒资 ID",
"FOLDERCODE"："媒资所属栏目 ID",
"NAME"："媒资名称",
"PORTAL_VER"："互动版本",
"SHOW_TYPE"："媒资类型",
"STBID"："智能卡号",
"TIME"："收藏时间"
}
```

2) 用户常看直播频道记录样例

```
{
"SID"："频道 serviceID",
"OPK"："区域码",
"STBID"："机顶盒卡号",
"L_CHANNEL_NAME"："频道名称",
"CNT"："观看次数",
}
```

2 数据清洗与数据预处理

原始数据集包含 2 份 csv 文件，特别注意的是，收藏记录这个 csv 文件是损坏的，需要用记事本手动打开删除文件最后一行的损坏值才能被 pandas 读入。收藏记录文件有 1171121 个观测值和 7 个特征，用户常看直播频道记录样例文件有 282900 个观测值和 5 个特征。在数据预处理部分中，我们的目标是通过做一些初步的统计分析、变量转换、NA 值处理等来熟悉数据集并做好准备。

2.1 “无用”属性

对于本聚类目标而言，我并不研究一个时间序列，用户收藏时间对于我聚类研究是没有意义的，这一列属性可以在预处理的时候删除。同时，用户收藏的 NAME 属性与 CODE 属性几乎是一一对应的关系，而且中文名称很难处理，这是一列冗余数据，所以 NAME 属性可以丢弃。同理的是在用户常看直播频道记录样例文件中，L_CHANNEL_NAME 属性如果算上缺失值和 OPK 属性也几乎一一对应，属于冗余数据，L_CHANNEL_NAME 属性也可以丢弃。

2.2 近零方差属性

接近零方差（NZV）指的是一些属性在整个数据集中每一条数据几乎取唯一值的情况。这类属性不仅有时是非信息性的，还可能破坏一些我们可能想要使用的数据挖掘方法。因此，删除这些属性列是一个好想法，特别是在处理具有大量变量的数据集时。在本问题的两个数据集中，有几个因子变量的方差接近于零。保留它们以后会产

生相当数量的虚拟变量，并增加计算的复杂性和对资源的需求。因此，我删除了所有 NZV 属性。默认情况下，如果样本中唯一值的百分比小于 10%，那么 R 语言中的 caret 包中的 nearZeroVar()函数将变量视为接近零的方差。在 python 中，我使用 pandas 的 value_count()函数，如果一个属性列出现最多次数的取值数量约占样本数量的 85%以上，我便选择删除这一列属性。通过在数据集上执行此操作，我删除了 2 个接近零方差的属性。下表展示是方差接近零的属性：

收藏记录：

PORTAL_VER	

用户常看直播频道记录样例文件：

OPK	

2.3 非数值属性值舍弃

要使用数值来计算距离等度量。为了允许在这些算法中使用分类属性，一种解决方法是使用一个热编码将这些分类属性转换成数值，但是这不是我在这个项目中采用的解决方案，原因很简单，无论是 CODE 属性，FOLDERCODE 属性，里面值的取值都太多了，热编码后的分类实在是太多，很多时候并没有聚类的价值，即使强行聚类，对计算量的开销也十分明显，所以最后我还是放弃了处理它们，决定将其舍弃

```
OTHE1000000001635517    22217
OTHE0000000001633666     7254
OTHE1000000001651838     6899
OTHE1000000001622517     6760
OTHE1000000001653678     6524
                        ...
OTHE0000000000642065        1
TWSX1520322922224145        1
OTHE0000000001505128        1
OTHE0000000001358830        1
OTHE1000000001887679        1
Name: CODE, Length: 12804, dtype: int64
MANU0000000000113860        155685
MANU0000000000114286         57067
MANU0000000000327801         44017
MANU0000000000114370         40176
MANU0000000000249853         37119
                             ...
5c048fc835db5ee92e8b456b         1
MANU0000000000313269             1
MANU0000000000249100             1
4d0086b32f2a241bd700bf9d         1
5b8dfa1b35db5e9a368b4684         1
Name: FOLDERCODE, Length: 4116, dtype: int64
```

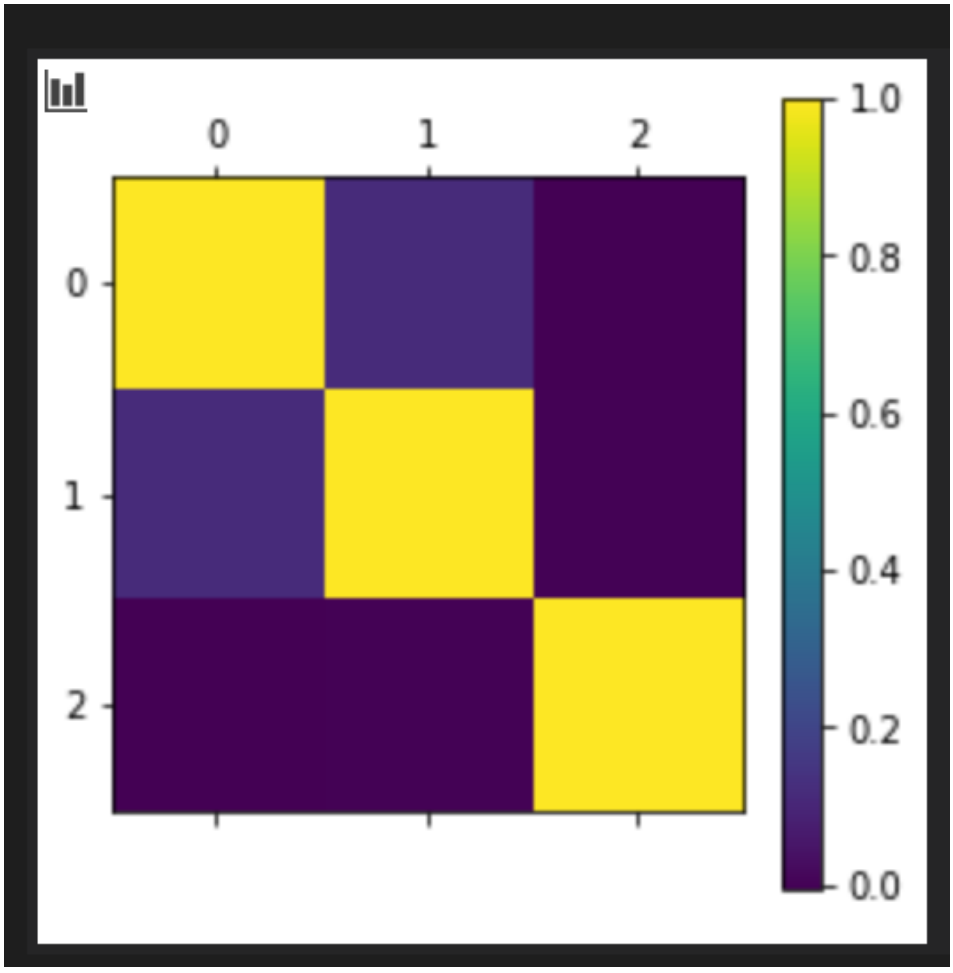
可以看到两个属性分别有 12804 个取值和 4116 个取值，对于一个分类属性而言，这个量太多了，很难找到可以聚类的指标，计算量很大但是效果平平，最终我决定将其舍弃。

2.4 数据表连接

本实验包含两个数据文件，为了方便聚类操作，考虑到 STBID 代表一个机顶盒的数据，用户收藏和观看记录之间的强关联关系，用 STBID 作为索引，内连接两个数据表，得到一个新表。同时为了节约内存空间，快速聚类的要求，只对连接后的数据表随机取样 1000 条数据，这样得到的数据表大小相对可控。（自然连接后的原始数据表有近 20 个 G，100 万数量级的数据确实太大太难找到特点）

2.5 相关系数

另一个需要探索的数值变量是它们之间的相关性。下面的热力图以绝对值显示了这些相关性，当颜色从深紫色变化到黄色时，相关性的绝对值从 0 增长到 1。统计了两两属性之间的相关系数大于 0.8 的组数为 0 组：一般来说，这些属性没有很强的相关性。



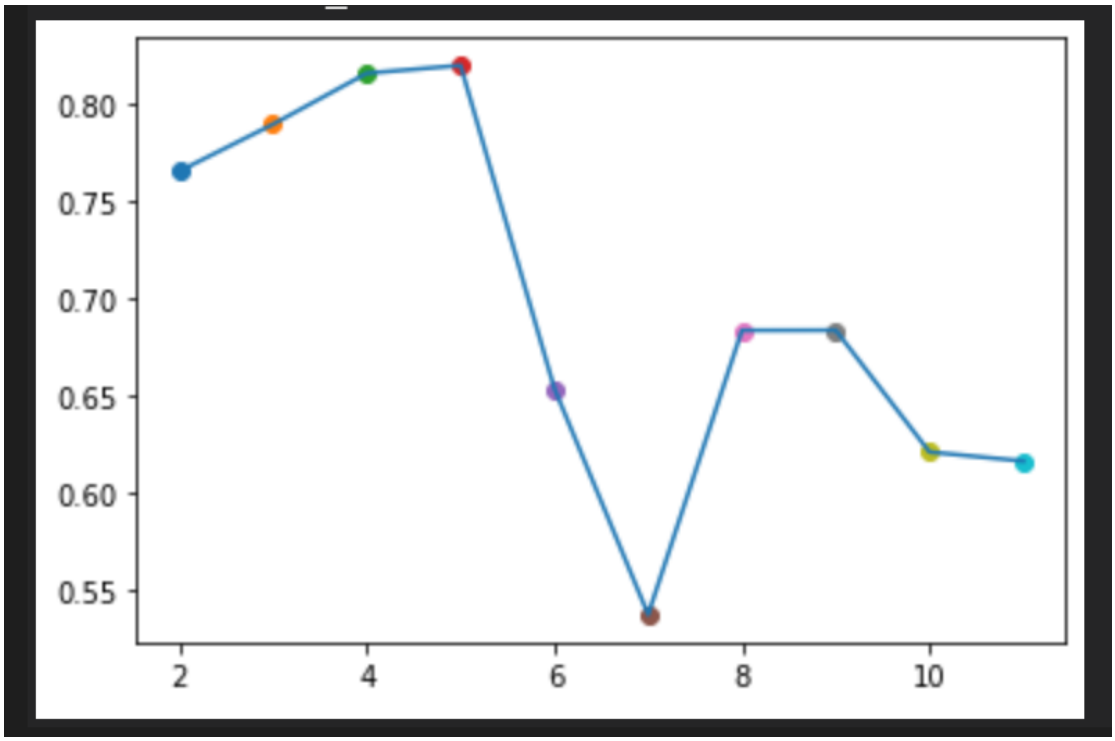
3 无监督学习

3.1 距离选择

由上述的数据预处理可知，我将数据中所有非数值值都删除了。很显然的，使用欧式距离是一个简单好用的选择，但是有些属性列数据分布跨度很大。因此，在欧式距离的基础上，我们对数据使用正态标准化，对标准化后的数据进行欧式距离计算、聚类等操作。

3.2 K-MEANS 聚类

由于没有进行 PCA 降维得到的结果，数据有三维不是很好可视化，无法看出数据被明显被分为几团，因此 K 中心数设置数并不知道。因此使用 2 到 11 遍历 K 中心数，计算设置中心数后的 SC 值。下图展示了轮廓系数随着 k 中心数变化而产生的变化。

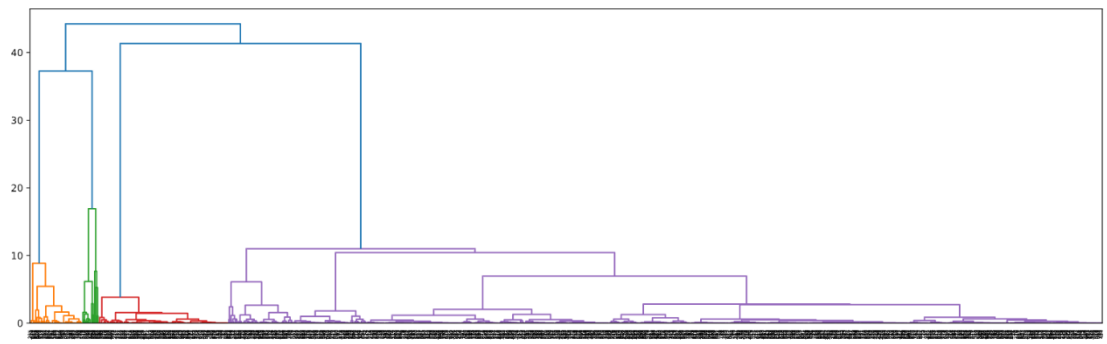


可见，在 K 中心值=5 的时候有轮廓系数的极大值 0.820，可以说是一个十分成功的聚类结果了。因此设置 k 中心值为 5，即聚类的类别数为 5。

3.3 层次聚类

3.3.1 树状图分析

在这一部分中，我通过使用这种时间层次聚类来处理数据集，这种聚类与 K-means 一样，将具有相似特征的数据点分组在一起，我使用的仍然是“euclidean”方法，即欧氏距离。我用 python 执行代码来计算树状图，得到下图：（原图见 picture 子文件夹下 hier.png 文件）



3.3.2 距离阈值切割分类

在构建树状图时，我决定在距离 15 处切割这棵树，得到 5 个聚类。计算轮廓系数，SC= 0.817,可见是一个成功的区分。

```
[14] import scipy.cluster.hierarchy as sch...

     The silhouette_score= 0.8172472346635709
```

4 实验结果讨论

两个聚类得到的聚类标签保存在 result2.csv 文件中

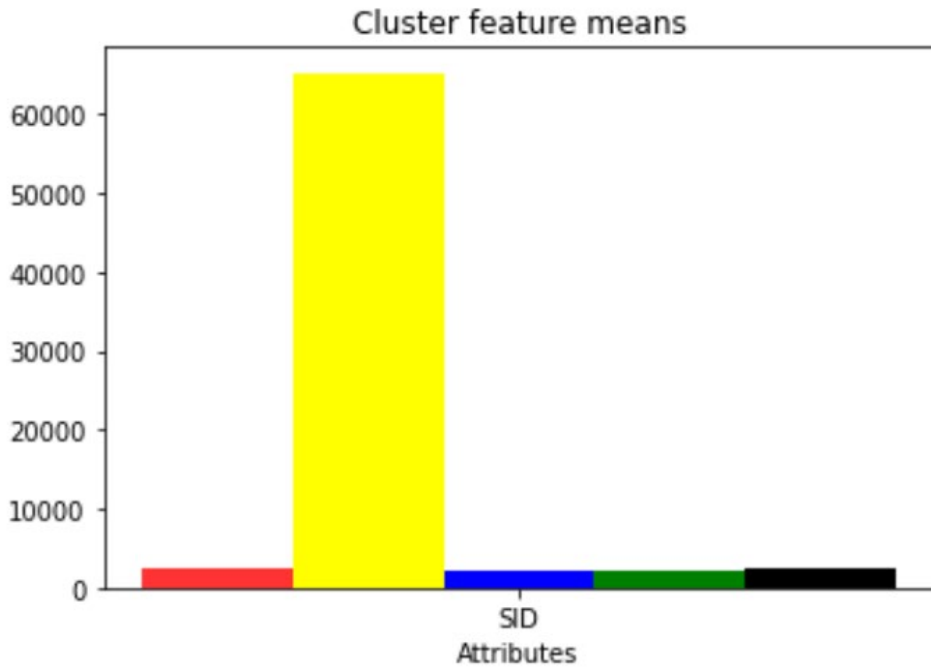
4.1 聚类算法开销比较

分析同实验 1，见 Q1 实验报告相应部分

4.2 聚类算法结果分析

从原始数据挑选出的三个数值属性，反应了一个用户对一个频道的观看次数，频道的类型和用户收藏节目的类型。从聚类分析得到的指标来看，这个是一个成功的聚类，SC 的值都在 0.8 左右。

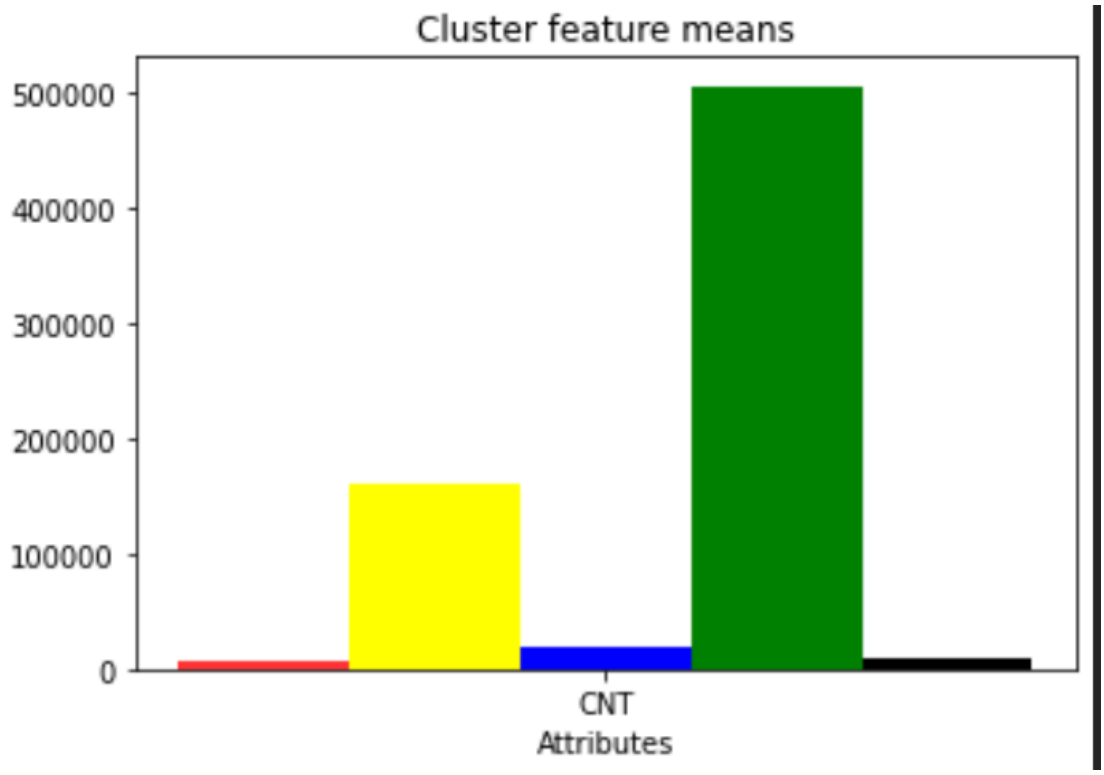
下列三张图是三个不同属性值的均值在各个聚类上的表现均值。根据这些结果我们可以分析:



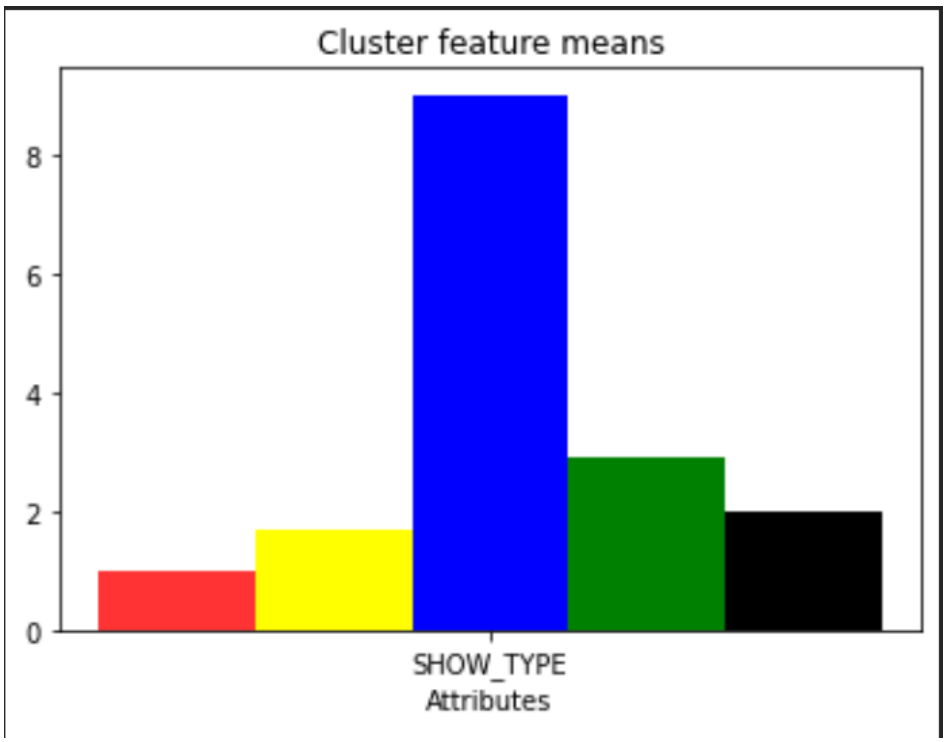
用户群体 2 普遍喜欢观看 SID 号偏大的频道，而其他用户群体对此并无明显偏好。查阅原始数据可知 SID 偏大的频道:

179	65325	1	8.35E+15	南平新闻	75	20210302	8
180	65325	1	8.35E+15	南平新闻	20	20210302	15
181	65325	1	8.35E+15	南平新闻	10	20210302	14
182	65325	1	8.35E+15	南平新闻	194	20210302	11
183	65325	1	8.35E+15	南平新闻	3091	20210302	7
184	65325	1	8.35E+15	南平新闻	7	20210302	5
185	65325	1	8.35E+15	南平新闻	21	20210302	17
186	65325	1	8.35E+15	南平新闻	4	20210302	18
187	65325	1	8.35E+15	南平新闻	1	20210302	8
188	65325	1	8.35E+15	南平新闻	28	20210302	14
189	65325	1	8.35E+15	南平新闻	11102	20210302	7
190	65325	1	8.35E+15	南平新闻	36	20210302	18
191	65325	1	8.35E+15	南平新闻	17	20210302	17
192	65325	1	8.35E+15	南平新闻	16	20210302	14
193	65324	1	8.35E+15	龙岩新闻	23772	20210302	14
194	65324	20	8.35E+15	龙岩新闻	6388	20210302	17
195	65324	1	8.35E+15	龙岩新闻	2	20210302	5
196	65324	20	8.35E+15	龙岩新闻	12	20210302	11
197	65324	20	8.35E+15	龙岩新闻	14426	20210302	13
198	65324	20	8.35E+15	龙岩新闻	5	20210302	18

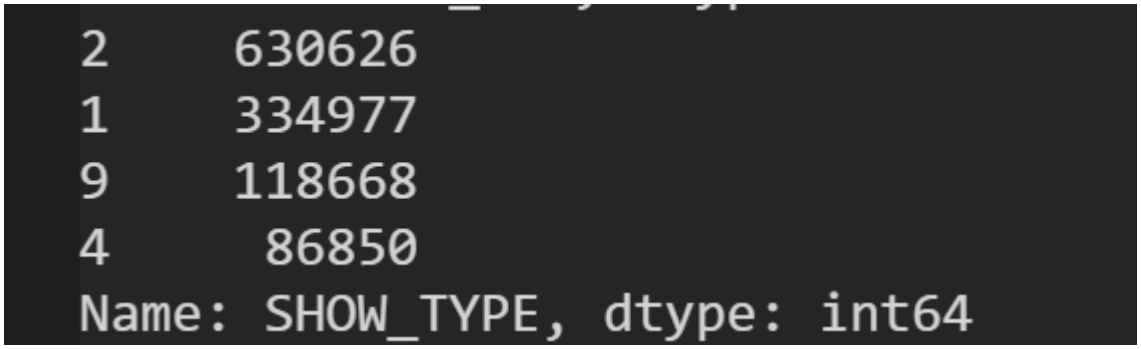
这些频道全是一些新闻频道，由此可以得出用户群 2 的画像是喜欢看新闻的用户，“新闻频道爱好者”，而其它聚类用户群并无此特征。



用户群体 4 在 CNT 这一项数值上的平均值明显超出很多，用户群体 2 在 CNT 这项数值上也偏高，其余用户群体的数值都差不多，CNT 是某一频道的观看次数。由此可知，用户群体 4 是某一频道的狂热爱好者，反复观看一个频道节目多次。而用户群体 2 虽然没有 4 狂热但是也是某一频道的热衷粉丝，由 SID 的结论可知，用户 2 群体是喜欢看新闻的一类用户。而这也很好解释为什么这些用户会有一定次数的重复观看记录，毕竟新闻是天天都有的嘛。



最后对于 SHOW_TYPE 这个属性，用户群体 3 一枝独秀。而由数据预处理时得到的信息（见下图），这一个属性只有 4 种取值，而用户群体 3 的这一项指标平均值高达 8+，说明用户群体 3 的用户几乎都收藏了类型 9 的节目。



查阅原始数据集可知（见下图），节目类型 9 多是综艺类节目，也有一定量的动画片节目被包含在其中。于是我又得到了用户群体 3 的用户画像是一些喜欢收藏综艺或者是动画片的人。

23	86492424		MANU000000000031	55b2ed6bl	龙岩风味	home	9	1	8.35E+15
24	86681726		MANU000000000031	85bdccd0(	泉州风味	home	9	1	8.59E+15
25	86682240	OTHE1000	MANU000000000024	de8d8b3d	熊出没	home	9	1	8.51E+15
26	86728751		MANU000000000031	59cb8b1d5	南平风味	home	9	1	8.51E+15
27	86733884	OTHE0000	MANU000000000031	2f315768f1	三明风味	home	9	1	8.51E+15
28	86991654		MANU000000000031	59cb8b1d5	南平风味	home	9	1	8.51E+15
29	86992292		MANU000000000031	5a56be3d(	三明风味	home	9	0	8.35E+15
30	87001585		MANU000000000031	85bdccd0(	泉州风味	home	9	1	8.6E+15
31	87090722		MANU000000000031	e54be05af	福州风味	home	9	1	8.59E+15
32	87120030		MANU000000000031	85bdccd0(	泉州风味	home	9	1	8.35E+15
33	87128472		MANU000000000031	5a56be3d(	三明风味	home	9	1	8.51E+15
34	88686119		MANU000000000026	7d203731e	高能少年E	home	9	1	8.35E+15
35	88717362	OTHE0000	MANU000000000031	7f5c95967(	漳州风味	home	9	0	8.51E+15
36	88759817		MANU000000000027	66b22a48f	奔跑吧2	home	9	1	8.51E+15
37	86199231		MANU000000000031	8321b6a5e	莆田风味	home	9	1	8.35E+15
38	86225251		MANU000000000031	0777ea37a	厦门风味	home	9	1	8.35E+15
39	86386864		MANU000000000031	0777ea37a	厦门风味	home	9	1	8.51E+15
40	93607171		MANU000000000031	d6242cad4	极限挑战5	home	9	0	8.51E+15
41	93691337		MANU000000000031	2c738aede	宁德风味	home	9	1	8.59E+15
42	95090438		MANU000000000031	e54be05af	福州风味	home	9	1	8.35E+15
43	95259620	OTHE0000	MANU000000000031	d1cb9400c	中餐厅3	home	9	0	8.51E+15
44	94804667		MANU000000000024	276b8a60(	小马宝莉	home	9	0	8.35E+15
45	95624047	OTHE0000	MANU000000000025	8c797c78c	快乐大本营	home	9	0	8.35E+15
46	95634074		MANU000000000031	0777ea37a	厦门风味	home	9	1	8.59E+15
47	95635845	OTHE0000	MANU000000000031	bdbb16f8(	厦门风味	home	9	1	8.59E+15
48	95636760		MANU000000000024	deae0c88c	珀利	home	9	1	8.6E+15

除此之外，用户群体 1 的聚类后属性均值只有 1。而由数据预处理时得到的信息，说明用户群体 1 的用户几乎都收藏了类型 1 的节目。

查阅原始数据集可知（见下图），节目类型 1 全是动画节目。于是我又得到了用户群体 1 的用户画像是一些喜欢收藏动画片的人。

6	1.04E+08	OTHE0000	MANU000000000031	e9a573e37	新大头儿子	home	1	1	8.6E+15	29-8月 -1
7	1.04E+08	OTHE0000	MANU000000000013	7b4fac4e6	追梦者	home	1	1	8.6E+15	29-8月 -1
8	1.05E+08	OTHE0000	MANU000000000011	d8b109fb8	精灵旅社2	home	1	0	8.6E+15	29-8月 -1
9	1.05E+08	OTHE0000	MANU000000000011	fdb3cf47ef	我知道你爱	home	1	0	8.6E+15	29-8月 -1
10	1.05E+08	OTHE0000	MANU000000000011	7bdeb406a	小猪佩奇过	home	1	0	8.35E+15	29-8月 -1
11	1.05E+08	OTHE0000	MANU000000000011	9b3831e5e	太空一号	home	1	1	8.6E+15	29-8月 -1
12	1.02E+08	OTHE0000	MANU000000000011	11d1a7be2	龙珠超：布	home	1	0	8.6E+15	27-8月 -1
13	1.02E+08	OTHE0000	MANU000000000024	da1ee5cb7	鼠来宝	home	1	0	8.6E+15	27-8月 -1
14	1.03E+08	OTHE0000	MANU000000000028	6fae249f4(	飞屋环游记	home	1	1	8.6E+15	03-11月-1
15	1.03E+08	OTHE0000	MANU000000000011	0065c88fc4	哥斯拉：怪	home	1	1	8.35E+15	27-8月 -1
16	1.03E+08	OTHE0000	MANU000000000011	73fc8fc49a	致青春：原	home	1	1	8.6E+15	28-8月 -1
17	1.03E+08	OTHE0000	MANU000000000024	7b2ea47e7	忍者神龟	home	1	1	8.6E+15	28-8月 -1
18	1.03E+08	OTHE0000	MANU000000000024	555b5cf7a	熊出没之雪	home	1	0	8.35E+15	28-8月 -1
19	1.03E+08	OTHE0000	MANU000000000011	a03e32f75	扫毒2：天	home	1	0	8.6E+15	28-8月 -1
20	1.03E+08	OTHE0000	MANU000000000024	805494824	爱宠大机密	home	1	0	8.6E+15	28-8月 -1
21	1.03E+08	OTHE0000	MANU000000000011	5392cedf9	一条狗的使	home	1	1	8.51E+15	28-8月 -1
22	1.03E+08	OTHE0000	MANU000000000011	5ee216595	史前巨鳄3	home	1	0	8.35E+15	28-8月 -1
23	1.04E+08	OTHE0000	MANU000000000025	4b1b9e24(	房间	home	1	0	8.35E+15	28-8月 -1