订单用户复购率的聚类分析

Summary

通过总体的分两组的人群，我们发现在双十一的复购率上有显明的差异。
那么我们现在对双十一前的这群订单用户，进行一个聚类，查看一下他们在双十一的一个复购率情况。

```
groupID = Flatten[orderIDList[weblog][[1 ;; -3]]] // DeleteDuplicates;
```

```
groupID // Length
```

```
5942
```

有六千个用户

```
ID双11 = orderIDList[weblog][[-2]];
```

```
inter = groupID ~ Intersection ~ ID双11;
% // Length
```

在双十一购买的有1832个用户

```
1
```

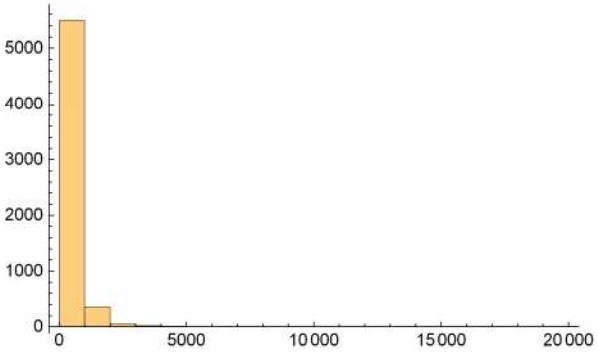
```
1832
```

```
inter // Length / groupID // Length // N
```

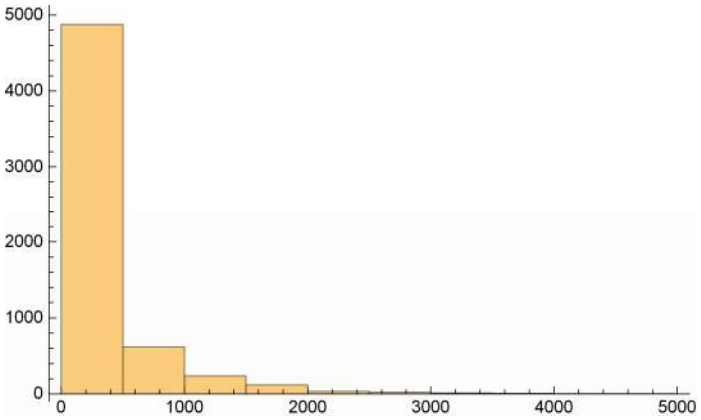
```
1
```

```
0.308314
```

复购率为0.308314，通过先前的分析，我们知道至少有一组的复购率在0.44，当然并不是说复购率越高就越好，以后还要考虑到成交金额等情况，如何给商家带来最大利润。那么，我们相信，通过聚类后，会有一群数量不少的人的人群的复购率最高，那么是否复购率最高的这群人，恰恰是先前花的钱最多的人群呢？我们暂时先使用成交金额来聚类。



成交金额少于5000的人数量分布



说明：具体实现过程中，聚类时遇到一些问题，不同的方法选择下，有不同的效果，受噪声的影响[聚类本身就有识别噪声的功能]，我觉得暂时效果一般，因为此例成交金额的情况使用分箱操作来替代，给价格分成6个区间，产生6个人群。

[0, 100][100 - 300][300 - 600][600, 1000][1000 - 3000], [3000 - ∞]

至于区间如何调整，使得各个区间有多少人，满足什么分布，则是一个专门可以优化的问题，比如哪一个价格区间的人数/成交金额/成本的一个综合情况达到了最优？我们可以使用一些条件，比如我想让每个区间的成交金额一样，来划分人数，或自然聚类划分等等

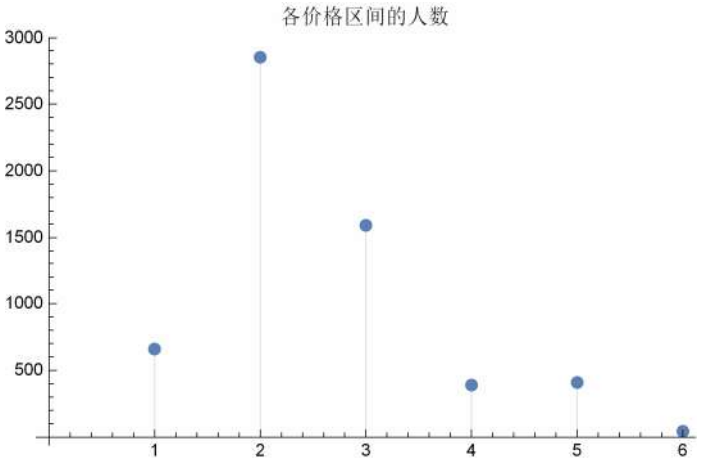
```
orderData = Table[
    (Rule[#[[1]], ToExpression@#[[2]]]) & /@ orderList[weblog][[i]], {i, 24}];

assoOrder[weblog] = Table[Association[
     (Rule[#[[1]], ToExpression@#[[2]]]) & /@ orderList[weblog][[i]]], {i, 24}];

assoGroupID = Merge[assoOrder[weblog][[1 ;; 22]], Total];

list = Reverse /@ Normal[assoGroupID];

d1 = BinCounts[list[[All, 1]], steps = {{0, 100, 300, 600, 1000, 3000, ∞}}]
```

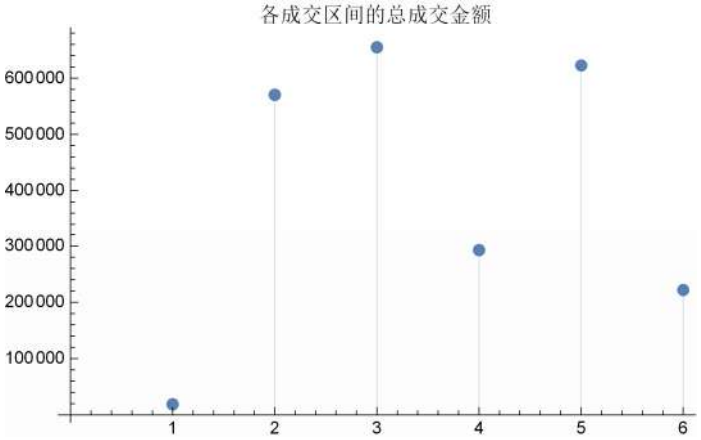
```
{659, 2850, 1593, 389, 408, 43}
```

在此实例中，100-300元消费金额的用户人群最多，而成交金额上来看，第二个区间的成交金额最大，并且虽然第五个区间的人数少，但是成交金额去并不低。广告投放时预算有限，当然要找效果最好的用户投广告。

```
intervals = Interval /@ Take[Partition[steps[[1]], 2, 1, 1], 6];
listBined =
  Table[Select[assoGroupID, IntervalMemberQ[intervals[[i]], #] &], {i, 6}];
listBoxed = BinLists[list[[All, 1]], steps];
payList = Total /@ listBoxed
```

```
{18 596.1, 570 106., 654 722., 292 504., 622 550., 221 411.}
```





子人群的在双十一的复购率情况

```
interList = Intersection[#, ID双11] & /@ (Keys /@ listBined);
```

子人群中各类的复购的人数

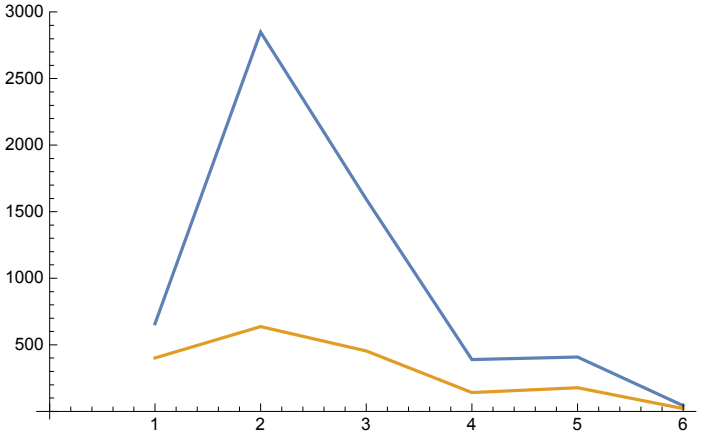
```
count1 = Length /@ interList
```

```
{402, 637, 454, 142, 178, 21}
```

原始购买的子人群的人数

```
{659, 2850, 1593, 389, 408, 43}
```

```
ListPlot[{d1, count1}, Joined → True, PlotRange → All]
```

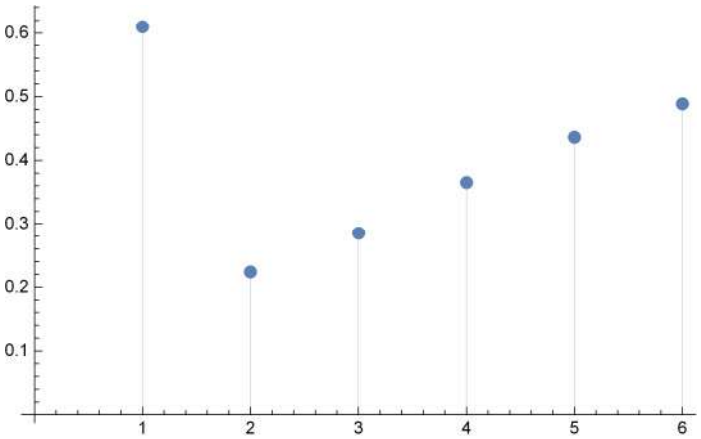


5类人群的双十一复购率

```
ratio = count1 / d1 // N
```

```
{0.610015, 0.223509, 0.284997, 0.365039, 0.436275, 0.488372}
```

```
ListPlot[ratio, Filling → Axis, PlotRange → All]
```



我们惊喜地发现，低端消费人群和高端消费人群的复购率是最高的，中端消费人群的复购率反而不高。消费金额低可能觉得再买一次的成本低，高端消费人群可能觉得花点钱无所谓。因为客户的消费占收入的比值不一样，中端消费用户可能更多会消费饱和

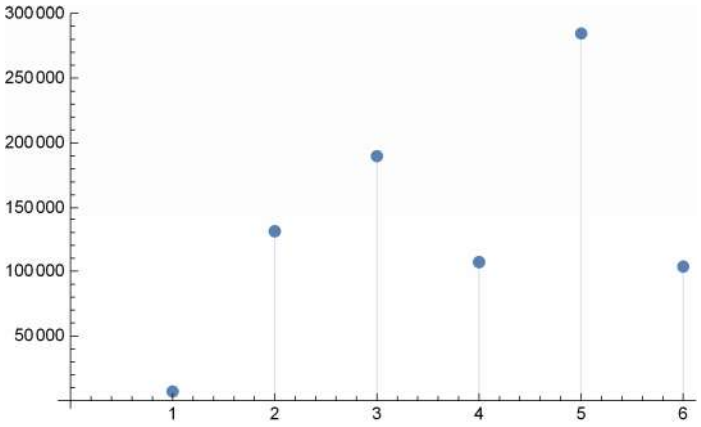
那么五类人群在消费金额上的表现如何呢？

果然，第1类人群只消费了7000块，第6类客户只消费了10万块，第5类客户的成交金额是最好的，从复购率来说，0.436也基本上达到了之前总体的复购率的一个高档的平均水平，效果不错。

```
payList = Table[Values[KeyTake[listBined[[i]], Key /@ ID双11]] // Total, {i, 6}]
```

```
{7038.95, 130 782., 189 880., 106 937., 284 542., 103 454.}
```

```
ListPlot[payList, Filling → Axis, PlotRange → All]
```



那么我们首先来看下第五类客户的一些浏览/购买车/收藏行为等指标的情况。

收藏【双十一的情况】

```
idCored = Keys[listBined[[5]]];
```

```
data[collect] = Partition[ReadList[pathFormat@
      "F:\\Documents\\Working\\统计分析\hdys\\data\\sub_collect_list.csv",
    "Record", "RecordSeparators" → {",", "\n"}], 7];
```

```
assoSub[collect] =
  GroupBy[Select[data[collect], #[[-1]] == "20141111" &], #[[6]] &];
```

```
ActNumberList[collect] = Values[Length /@ assoSub[collect]];
```

双11的平均收藏次数非常少，只有2，收藏量在14898，后续我们会给出五类人的一个收藏，购物车情况。
因为双十一的宝贝，在活动过后，有许多会失效，收藏没有特别的意义。

```
Mean@ActNumberList[collect] // N // Round
```

1

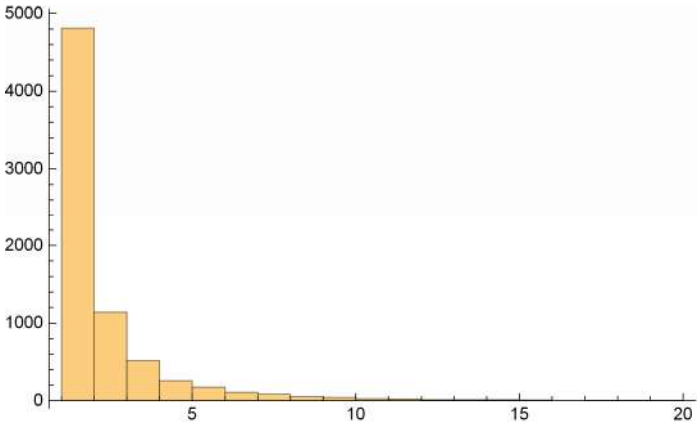
2

```
Total@ActNumberList[collect] // N // Round
```

1

14 898

```
Histogram@Select[ActNumberList[collect], # < 20 &]
```



购物车【双十一】

```
data[cart] = Partition[ReadList[pathFormat@
      "F:\\Documents\\Working\\统计分析\hdys\\data\\sub_cart_list.csv",
     "Record", "RecordSeparators" → {",", "\n"}], 7];

assoSub[cart] = GroupBy[Select[data[cart], #[[-1]] == "20141111" &], #[[6]] &];

ActNumberList[cart] = Values[Length /@ assoSub[cart]];
```

双11的购物车平均收藏次数有3，但是量明显非常大，有126128，购物车数/收藏数之比为

```
Mean@ActNumberList[cart] // N // Round
```

1

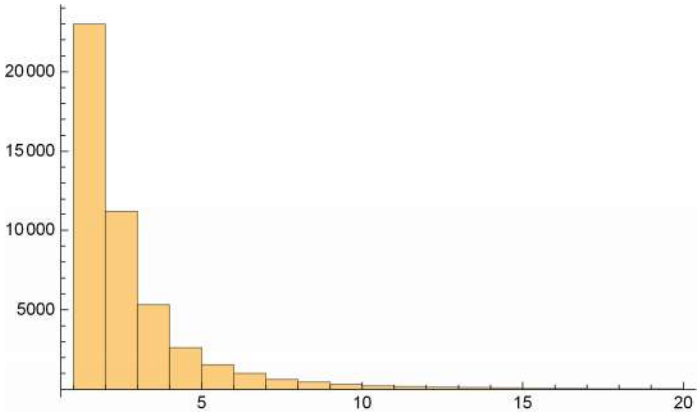
```
3
```

```
Total@ActNumberList[cart] // N // Round
```

1

```
126128
```

```
Histogram@Select[ActNumberList[cart], # < 20 &]
```



收藏【双十一前22天】

```
idCored = Keys[listBined[[5]]];
```

```
data[collect] = Partition[ReadList[pathFormat@
      "F:\\Documents\\Working\\统计分析\hdys\\data\\sub_collect_list.csv",
     "Record", "RecordSeparators" → {",", "\n"}], 7];
```

```
assoSub[collect] = GroupBy[Select[data[collect],
    #[[-1]] != "20141111" || #[[-1]] != "20141112" &], #[[6]] &];
```

```
ActNumberList[collect] = Values[Length /@ assoSub[collect]];
```

双11前的平均收藏数较有3，收藏量在153777，对比双十一数据明显。

```
Mean@ActNumberList[collect] // N // Round
```

1

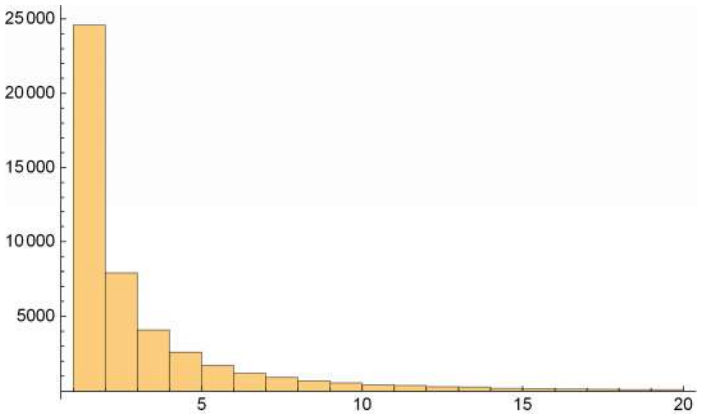
3

```
Total@ActNumberList[collect] // N // Round
```

1

153 777

```
Histogram@Select[ActNumberList[collect], # < 20 &]
```



购物车【双十一前22天】

```
data[cart] = Partition[ReadList[pathFormat@
      "F:\\Documents\\Working\\统计分析\hdys\\data\\sub_cart_list.csv",
     "Record", "RecordSeparators" → {",", "\n"}], 7];
```

```
assoSub[cart] = GroupBy[
   Select[data[cart], #[[-1]] != "20141111" || #[[-1]] != "20141112" &], #[[1]] &];
```

```
ActNumberList[cart] = Values[Length /@ assoSub[cart]];
```

双11前的平均购物车数量相对较更多，有6，但是总体的购物车数量还是非常大的，529711，对比双十一数据明显。

```
Mean@ActNumberList[cart] // N // Round
```

1

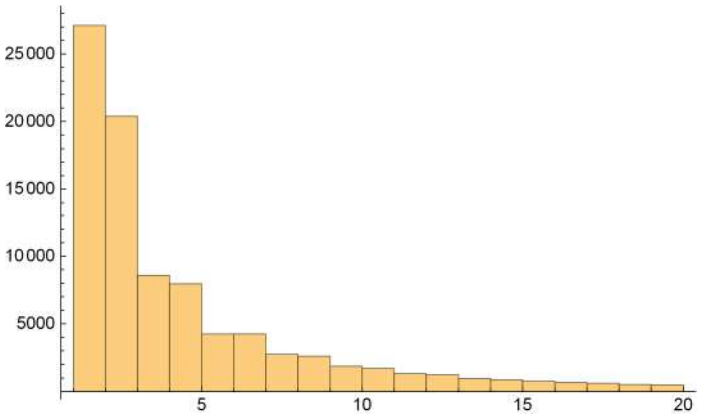
6

```
Total@ActNumberList[cart] // N // Round
```

1

529 711

```
Histogram@Select[ActNumberList[cart], # < 20 &]
```



五类人群的收藏和购物车情况对比

双十一

```
data[cart] = Partition[ReadList[pathFormat@
      "F:\\Documents\\Working\\统计分析\hdys\\data\\sub_cart_list.csv",
     "Record", "RecordSeparators" → {",", "\n"}], 7];
```

```
assoSub[cart] = GroupBy[Select[data[cart], #[[-1]] == "20141111" &], #[[1]] &];
```

```
numberList[cart] = Table[
   Values[Length /@ KeyTake[assoSub[cart], Key /@ Keys[listBined[[i]]]]], {i, 6}];
```

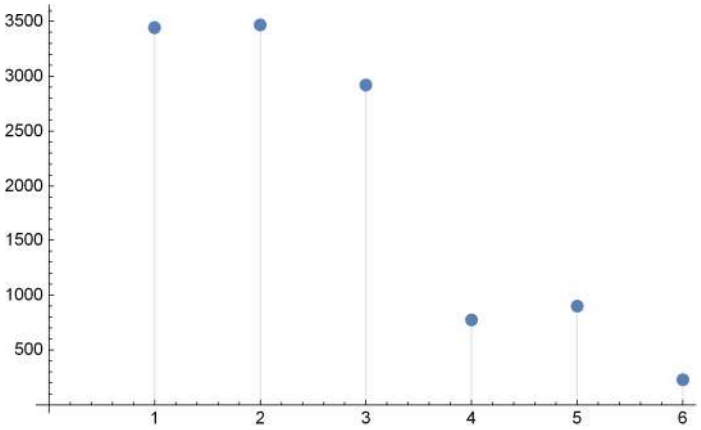
```
n1 = Mean /@ numberList[cart] // N // Round
```

```
{8, 4, 5, 5, 5, 10}
```

```
n2 = Total /@ numberList[cart] // N // Round
```

```
{3442, 3466, 2919, 772, 899, 228}
```

```
ListPlot[n2, Filling → Axis]
```



我们看到，添加购物车数上，第二类最多，第一类次之，第六类高端人群数量少，但是平均，第六类是最高的，有10，第1类有8。

```
n1 = Mean /@ numberList[cart] // N // Round
```

```
{8, 4, 5, 5, 5, 10}
```

双十一前购物车

```
data[cart] = Partition[ReadList[pathFormat@
      "F:\\Documents\\Working\\统计分析\hdys\\data\\sub_cart_list.csv",
     "Record", "RecordSeparators" → {",", "\n"}], 7];
```

```
assoSub[cart] = GroupBy[
   Select[data[cart], #[[-1]] != "20141111" || #[[-1]] != "20141112" &], #[[1]] &];
```

```
numberList[cart] = Table[
   Values[Length /@ KeyTake[assoSub[cart], Key /@ Keys[listBined[[i]]]]], {i, 6}];
```

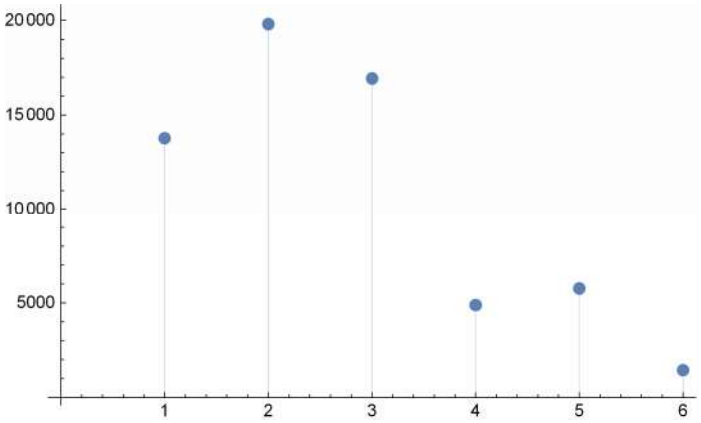
```
n1 = Mean /@ numberList[cart] // N // Round
```

```
{23, 9, 13, 15, 18, 35}
```

```
n2 = Total /@ numberList[cart] // N // Round
```

```
{13 777, 19 810, 16 933, 4881, 5757, 1439}
```

```
ListPlot[n2, Filling → Axis]
```



第五类双十一前和双十一时有差异。

其他---

挑选出一群人在成交/购物/收藏时表现明显的人群，可以跟踪他们的行为和属性，去把握好的特征，看有什么特点，人群透视，然后可以进行人群放大，重点投放，也可以修改产生人群的模型等。

如何挑选出最优质的一群客户，如利润最大，成交情况最好及其他各种指标等等是一个大问题与好问题，是一个优化问题。

目前的情况，我们可以手动分成，或自动分成一些类，查看子类的一个情况

举例来说，会有这样的一个问题是，比如均匀分成了四类，[0,25][26,50][51-75][76-100]虽然可能我们发现[51-75]的效果是相对其他类来说最好的，但是可能[60-80]是一个最优的类。

10月20日购买的人群，在未来22天的复购率，在双十一的复购人数占比，金额占比等

地域分析，通过聚类，我们可以发现，哪个人群的地域情况，这些地域情况可以重点投放，人群的地

域属性在周期内的变化特点等。

```
cells =
  Select[NotebookRead@Cells[], #[[2]] != "Input" && FreeQ[#, (CellTags → "1")] &];
```