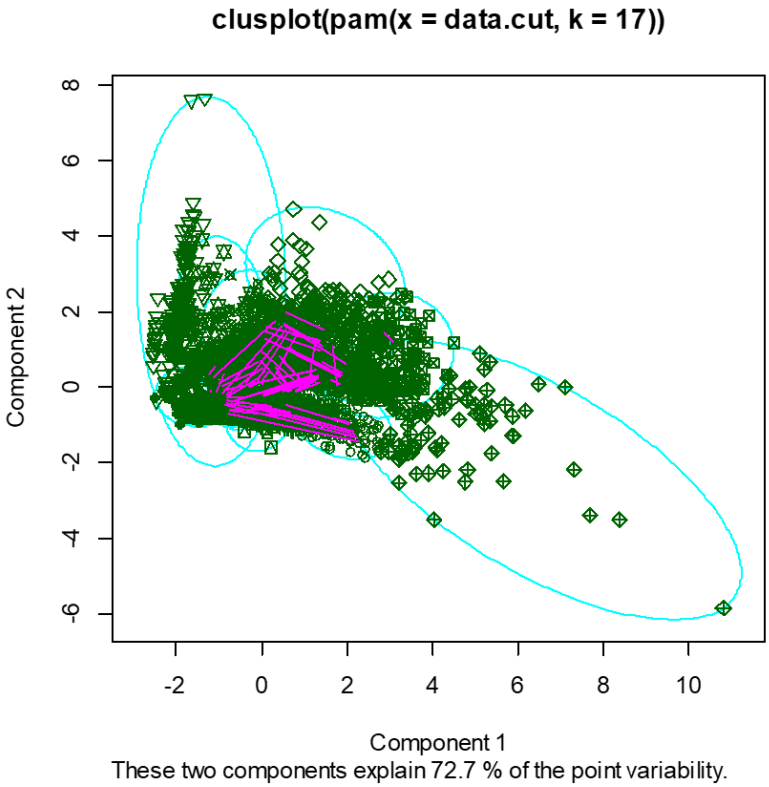
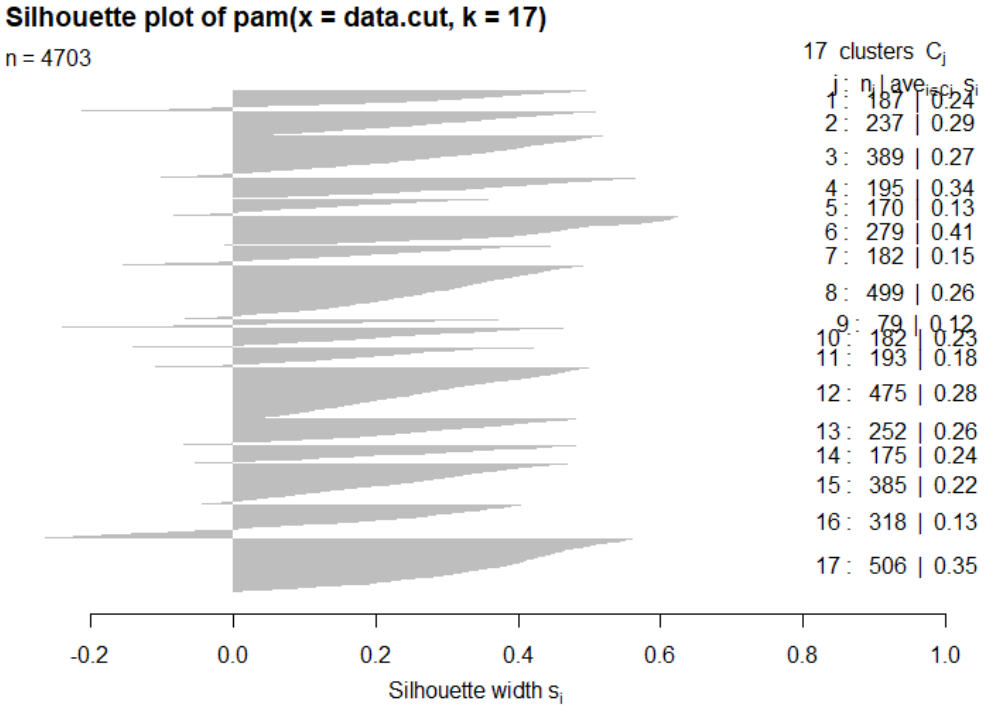
PAM

1. Firstly, the data obtained by SAS after dimensionality reduction were processed to obtain the data used for K-methods analysis.

Factor1	Factor2	Factor3	Factor4
1.0957	-0.4099	-0.4004	1.3764
1.2496	1.0433	1.3544	-0.4226
0.6065	0.7071	-0.4714	0.4641
1.1946	0.8062	-0.2757	0.9940
2.1832	0.8749	1.4010	0.7490
0.8249	-0.3740	1.6784	-0.2971
-1.2824	-3.5999	0.5087	-0.3153
1.2996	0.8091	1.3359	0.4076
1.8321	0.2421	0.3473	1.6793
1.2273	1.0194	1.3372	0.0425
1.0657	0.8317	1.2323	-0.2408
0.3239	0.8930	-0.1936	0.0883

2. According to the clustering results obtained by SAS, the value of the center point was set to 17 for further comparison, and k- medoids clustering was carried out for the data.

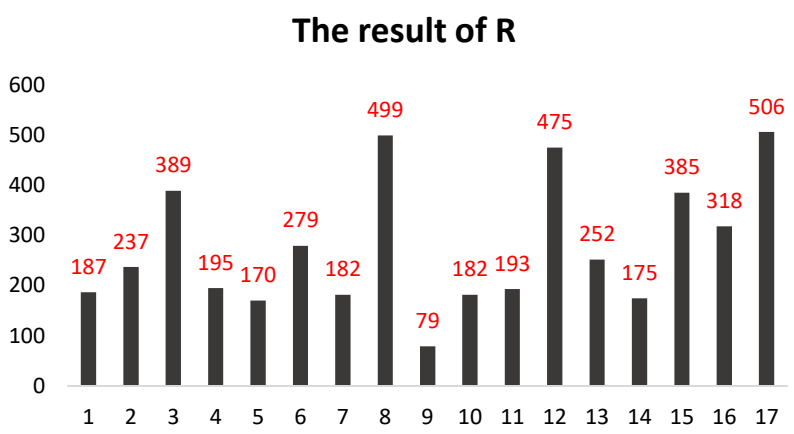


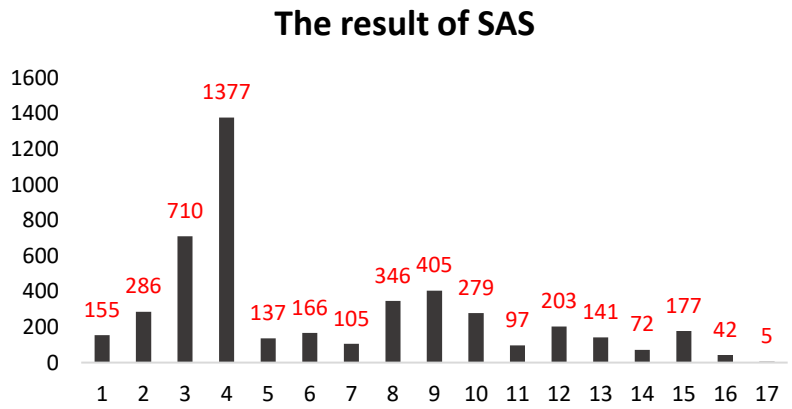


3. Get the center point of the 17 classes.

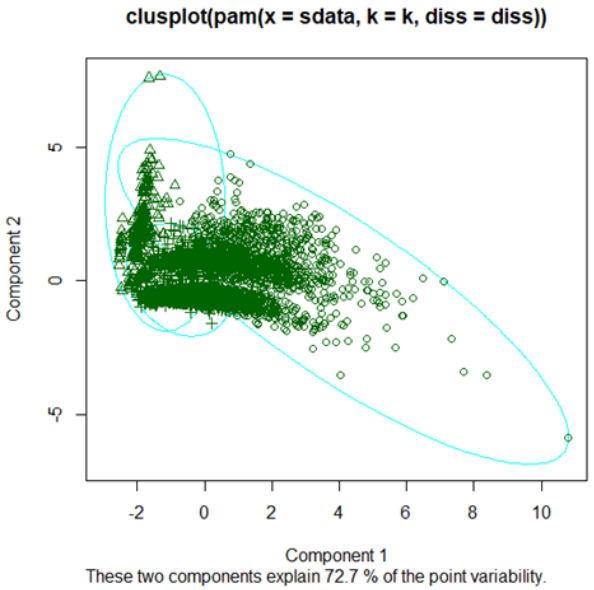
medoids	Factor1	Factor2	Factor3	Factor4
1	1.2060	0.4051	-0.5400	1.4712
2	0.5180	0.6977	1.2849	-0.3136
3	0.6786	0.6375	-0.3802	0.4131
4	1.3940	0.6868	1.2441	0.5858
5	1.1926	-0.2461	2.1286	0.1435
6	-1.2824	-3.3049	0.3029	-0.3157
7	2.0216	-0.3029	0.9430	1.7573
8	0.1448	0.5968	-0.4572	-0.1158
9	2.7353	0.0884	0.1369	4.4462
10	0.5974	-0.1495	0.5975	0.3312
11	-0.9186	-0.4798	0.6543	-0.6041
12	-0.2508	0.3309	-0.6604	-0.3676
13	-0.3646	0.7070	1.3149	-0.8808
14	0.0367	-0.4502	-0.7734	0.2565
15	-0.6199	0.4370	-0.4955	-0.5281
16	-1.1064	-0.4466	-1.0355	-0.4853
17	-0.9928	0.0911	-0.8110	-0.5999

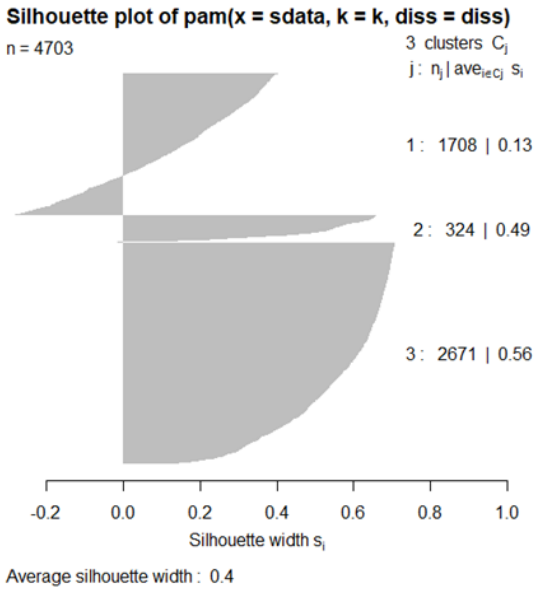
4. The results obtained by R PAM classification were compared with those obtained by SAS, As can be seen from the following figure, the classification results of R and SAS are quite different.





5. Because the classification is too detailed, it is not conducive to further study. Therefore, PAM is optimized and FCP is adopted to select the optimal classification number. The optimal number of categories selected by fcp is three, and it can be seen from the drawing that the effect of three categories is better.





RPART

1.Firstly, according to the dimensionality reduction results obtained by sas, the training set and test set are selected. The data**4** of SALES=NA are the test set and the rest are the training set.

SALES	Factor1	Factor2	Factor3	Factor4
4	1.0957	-0.4099	-0.4004	1.3764
NA	1.2496	1.0433	1.3544	-0.4226
1	0.6065	0.7071	-0.4714	0.4641
57	1.1946	0.8062	-0.2757	0.9940
7	2.1832	0.8749	1.4010	0.7490
17	0.8249	-0.3740	1.6784	-0.2971
809	-1.2824	-3.5999	0.5087	-0.3153
NA	1.2996	0.8091	1.3359	0.4076
790	1.8321	0.2421	0.3473	1.6793
586	1.2273	1.0194	1.3372	0.0425
NA	1.0657	0.8317	1.2323	-0.2408
19	0.3239	0.8930	-0.1936	0.0883
15	-1.2824	-4.2787	1.8498	-0.3385
6	1.2529	0.7131	0.7362	1.2431
132	-1.2824	-3.7545	0.4273	-0.3099
28	0.2078	0.6722	-0.3491	-0.2253
224	0.3850	1.0210	0.0654	-0.0052
152	0.3771	0.8352	-0.0049	-0.0730
15	0.5326	1.0479	0.0198	-0.3310
6	0.3300	0.8581	1.1419	-0.5649

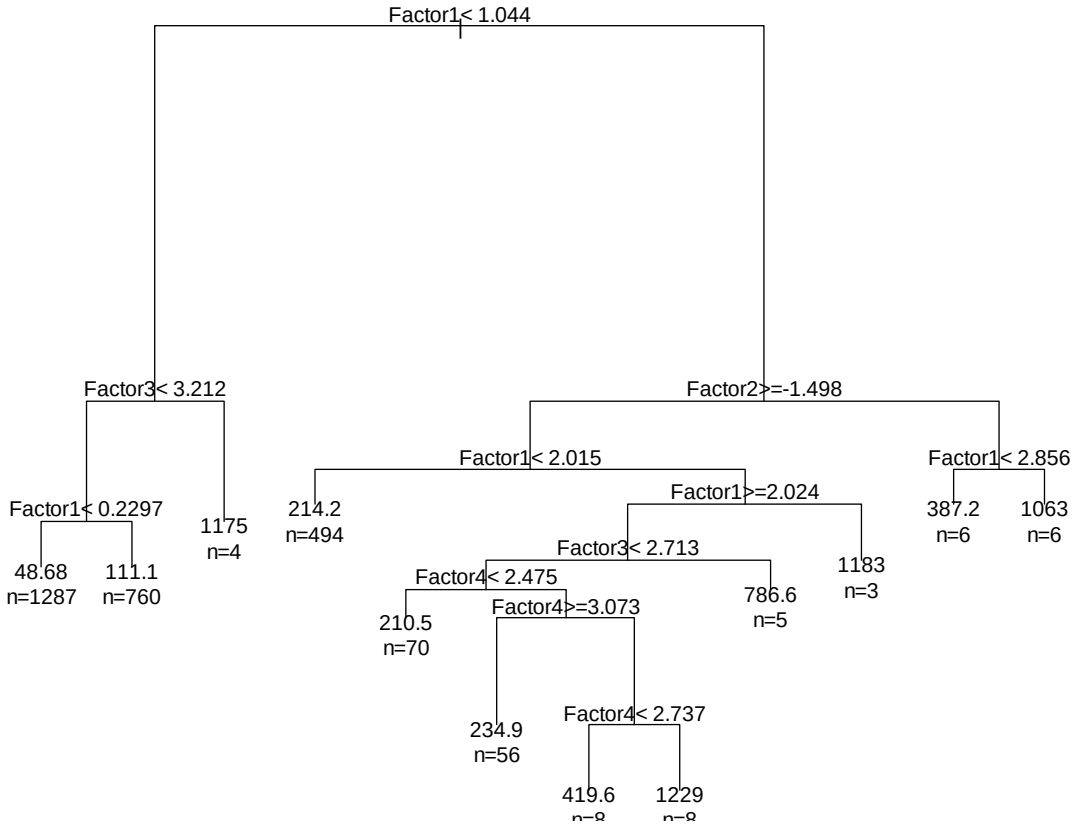
2. The obtained training set and test set length

item	length
training set	2707
test set	1996

3. The results of the tree method obtained from the test set are as follows

```
1) root 2707 130556000.0    116.32880
    2) Factor1< 1.043504 2051    40479120.0      74.02243
      4) Factor3< 3.212448 2047    27732720.0      71.87103
        8) Factor1< 0.2296959 1287    11760040.0      48.68454 *
        9) Factor1>=0.2296959 760    14109080.0    111.13550 *
      5) Factor3>=3.212448 4      7888314.0 1175.00000 *
    3) Factor1>=1.043504 656    74928670.0    248.60060
      6) Factor2>=-1.498155 644    64931910.0    239.72050
       12) Factor1< 2.014721 494    27123700.0    214.20450 *
       13) Factor1>=2.014721 150    36427360.0    323.75330
         26) Factor1>=2.024357 147    32881060.0    306.21090
           52) Factor3< 2.713288 142    30069120.0    289.29580
            104) Factor4< 2.475297 70      4373779.0    210.48570 *
            105) Factor4>=2.475297 72    24837870.0    365.91670
              210) Factor4>=3.073342 56      5100261.0    234.89290 *
```

```
      211) Factor4< 3.073342 16    15411470.0    824.50000
        422) Factor4< 2.736663 8       721995.9    419.62500 *
        423) Factor4>=2.736663 8    12066690.0 1229.37500 *
     53) Factor3>=2.713288 5      1617441.0    786.60000 *
   27) Factor1< 2.024357 3      1284435.0 1183.33300 *
 7) Factor2< -1.498155 12      7220588.0    725.16670
  14) Factor1< 2.855771 6        512750.8    387.16670 *
  15) Factor1>=2.855771 6      5336909.0 1063.16700 *
```



4. Select the optimal number of nodes, CP value was selected according to the principle of minimum xerror, which was not applicable at this time, because too many branches were cut from the decision tree, resulting in inaccurate final prediction results. Moreover, due to the small number of CP values, manual deletion was adopted to select the optimal CP value as 0.01050069

item	CP	nsplit	Rel error	xerror	xstd
1	0.11602842	0	1.0000000	1.0008277	0.1690388
2	0.03721071	1	0.8839716	0.8985658	0.1615232
3	0.02126418	2	0.8467609	0.9370605	0.1627320
4	0.01427434	3	0.8254967	0.9455882	0.1624032
5	0.01395082	4	0.8112223	0.9408809	0.1624884
6	0.01050069	10	0.7143780	0.9490305	0.1627981
7	0.01000000	11	0.7038773	0.9519849	0.1628204

5. Predict sales according to the optimal CP value and make it into a categorical

SALES	Factor1	Factor2	Factor3	Factor4	level
156	-1.2824	-2.1311	3.0462	-1.3295	3
192	1.4937	0.9005	-0.1112	0.6364	3
65	1.9688	0.9642	1.4695	0.7248	2
26	0.3780	0.7264	-0.0770	-0.0163	1
74	0.5180	0.6977	1.2849	-0.3136	2
1	0.4147	0.5612	-0.4775	-0.0839	1
738	1.1555	0.5503	-0.3756	1.2269	3
6	0.4190	0.6199	-0.2977	0.0207	1
8	0.3317	-0.3685	0.6162	0.0342	1
32	1.2323	0.8899	-0.1468	0.9761	1
3	2.7721	0.0543	2.1291	1.5887	1
16	0.6011	0.5854	1.1163	-0.3800	1
4	-0.0443	0.6043	-0.2336	-0.4322	1
24	-1.3879	-1.3415	0.8960	-1.1500	1
12	-0.6091	0.4108	-0.4971	-0.4391	1
214.2045	1.2496	1.0433	1.3544	-0.4226	3
48.68454	-1.2824	-4.3541	-1.1952	-0.2525	1
111.1355	0.3766	0.9557	1.5958	-0.9392	3