Задание:
Нужно выполнить систематизацию наблюдений по нескольким признакам (от 5 до 10) при помощи методов главных компонент и кластерного анализа. Важно выяснить, насколько можно доверять восстановлению данных по первым двум главным компонентам. В кластерном анализе выполнить кластеризацию индивидов, признаков и выделить оптимальное число кластеров по индивидам. В качестве рабочего материала можно взять кардиологические данные финских алкоголиков. Использование других данных не возбраняется.

Метод главных компонент

	HR.1	SBP.1	DBP.1	MBP.1	SV.1	CO.1	SI.1	CI.1	TPR.1
HR.1	1.000	0.144	0.066	0.285	-0.478	-0.055	-0.506	-0.056	0.095
SBP.1	0.144	1.000	0.626	0.527	0.013	0.206	0.032	0.200	-0.036
DBP.1	0.066	0.626	1.000	0.362	-0.355	-0.273	-0.290	-0.263	0.362
MBP.1	0.285	0.527	0.362	1.000	-0.105	0.075	-0.052	0.117	-0.047
SV.1	-0.478	0.013	-0.355	-0.105	1.000	0.850	0.955	0.826	-0.452
CO.1	-0.055	0.206	-0.273	0.075	0.850	1.000	0.844	0.978	-0.527
SI.1	-0.506	0.032	-0.290	-0.052	0.955	0.844	1.000	0.864	-0.437
CI.1	-0.056	0.200	-0.263	0.117	0.826	0.978	0.864	1.000	-0.502
TPR.1	0.095	-0.036	0.362	-0.047	-0.452	-0.527	-0.437	-0.502	1.000

Матрица корелляций (данные центрированы и усреднены, поэтому совпадает с ковариацонной). По ней можно судить о независимости признаков, например, HR.1 и SBP.1 (0.144). Или же, наоборот, линейной зависимости: SI.1 и SV.1 (0.955).

	Eigen.values	Percent
1	4.213	46.808
2	2.105	70.194
3	1.185	83.359
4	0.665	90.745
5	0.509	96.402
6	0.222	98.871
7	0.064	99.586
8	0.031	99.930
9	0.006	100.000

Представлены собственные числа корелляционной матрицы в порядке убывания вклада в общую дисперсию

```
[1] 9 9
```

Убедились в том, что сумма собственных чисел совпадает с суммарной дисперсией нормированных признаков

```
 [1] 4.21270569 2.10478551 1.18479052 0.66480073 0.50907856 0.22226553 0.06433614 0.03091845 0.00631886
[10] 4.21270569 2.10478551 1.18479052 0.66480073 0.50907856 0.22226553 0.06433614 0.03091845 0.00631886
```

Убедились, что дисперсии главных компонент совпадают с собственными числами корреляционной матрицы

[1] "Factors"

	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	0.196	1.311	0.274	0.391	0.488	-0.544	-0.574	0.568	-0.740
2	2.749	0.378	0.300	-0.357	0.080	1.569	1.123	-0.256	1.108
3	-0.377	-0.207	0.763	-0.119	-0.176	-0.484	0.205	0.263	-0.117
4	-0.499	2.140	-2.264	-1.355	-4.440	0.231	0.009	-0.167	-0.253
5	-1.146	0.082	-0.097	-0.522	0.643	1.350	-0.018	0.401	0.332
6	-0.192	-1.595	1.070	-0.294	-0.899	0.391	0.813	-0.099	1.047
7	-0.653	0.418	1.522	-0.323	-0.246	-1.187	1.188	0.405	1.673
8	-0.268	0.962	0.453	0.434	0.269	0.743	0.398	1.521	0.844
9	1.260	-2.092	0.841	-0.518	-1.090	0.846	-0.430	0.801	-1.138
10	-0.548	0.982	1.795	-0.555	-0.040	-0.099	-1.031	-1.129	-1.621
11	-1.697	-0.027	0.896	-1.898	0.633	1.068	0.710	-1.773	1.459
12	1.699	1.017	0.445	0.048	0.251	-0.100	-0.200	-0.178	-0.208
13	1.580	1.099	0.993	-0.269	0.069	-1.417	-0.612	0.856	-0.673
14	-0.838	-0.993	-2.011	-0.167	1.043	-0.401	0.046	0.604	-0.816
15	0.603	-1.280	0.090	-0.004	-0.350	-0.663	-4.049	-0.760	1.013
16	-0.670	1.035	1.792	-0.579	-0.004	-0.637	-0.356	-1.102	-0.031
17	-0.441	0.443	0.329	-0.068	0.336	-1.467	-0.120	0.426	-0.815
18	-0.821	0.731	-1.013	-0.236	1.250	0.417	-1.139	0.571	-0.593
19	1.393	0.885	-1.423	0.093	1.291	-0.923	0.879	-3.184	0.820
20	0.288	0.914	-0.772	0.642	0.781	-0.271	0.290	0.245	-0.391
21	-0.254	-0.426	-0.007	0.005	0.122	0.049	-0.884	-0.426	-0.917
22	0.483	-0.658	-0.889	0.453	0.260	-0.397	0.424	-0.152	-0.389
23	-0.424	1.175	-0.050	0.393	0.649	0.958	-0.350	1.227	0.506
24	0.908	0.039	0.440	0.225	-0.213	0.272	0.727	1.705	0.418
25	0.840	-0.791	-0.276	0.223	-0.090	-0.631	0.473	0.515	-0.421
26	-1.628	-0.116	-0.959	-1.455	1.334	0.315	0.019	1.136	-0.227
27	0.071	-0.569	0.115	0.160	-0.178	-1.171	1.001	1.429	0.441
28	-0.063	-1.123	-0.156	0.114	-0.154	1.054	-0.437	-0.802	-0.426
29	0.285	-0.156	0.151	0.494	-0.029	2.900	-0.587	-0.253	0.207
30	-0.823	-1.783	0.147	-0.678	-0.727	-1.958	1.151	-0.601	-0.267
31	-1.729	0.187	0.497	4.616	-1.076	0.143	0.174	-0.774	0.322
32	0.360	-0.104	-1.443	0.604	0.768	-0.286	0.574	-0.622	0.157
33	0.192	-0.991	-1.453	0.207	-0.356	-0.658	-0.943	0.451	2.498
34	0.165	-0.888	-0.103	0.297	-0.197	0.985	1.526	-0.844	-2.801

Матрица факторов

	X1	X2	X3	X4	X5	X6	X7	X8	X9
HR.1	-0.354	-0.351	0.792	-0.312	-0.119	-0.057	0.065	-0.060	0.007
SBP.1	0.030	-0.890	-0.196	0.051	-0.257	0.314	-0.009	-0.019	0.002
DBP.1	-0.423	-0.671	-0.476	-0.045	-0.199	-0.319	0.027	0.002	-0.002
MBP.1	-0.056	-0.784	0.157	0.223	0.554	-0.025	0.009	0.012	0.001
SV.1	0.953	0.047	-0.193	-0.079	0.059	0.030	0.201	-0.013	-0.025
CO.1	0.926	-0.238	0.157	-0.196	-0.076	-0.031	0.002	0.118	0.038
SI.1	0.953	-0.002	-0.236	-0.075	0.099	-0.054	-0.051	-0.112	0.040
CI.1	0.920	-0.254	0.151	-0.207	-0.015	-0.060	-0.126	-0.003	-0.052
TPR.1	-0.617	0.023	-0.355	-0.648	0.251	0.097	-0.010	0.011	0.001

Матрица факторных нагрузок. По ней можно судить о весомости факторов. Так, фактор X1 имеет высокую корреляционную связь с большинством из выбранных признаков. Его необходимо интерпретировать.

	lambda	EI
1	4.213	4.213
2	2.105	2.105
3	1.185	1.185
4	0.665	0.665
5	0.509	0.509
6	0.222	0.222
7	0.064	0.064
8	0.031	0.031
9	0.006	0.006

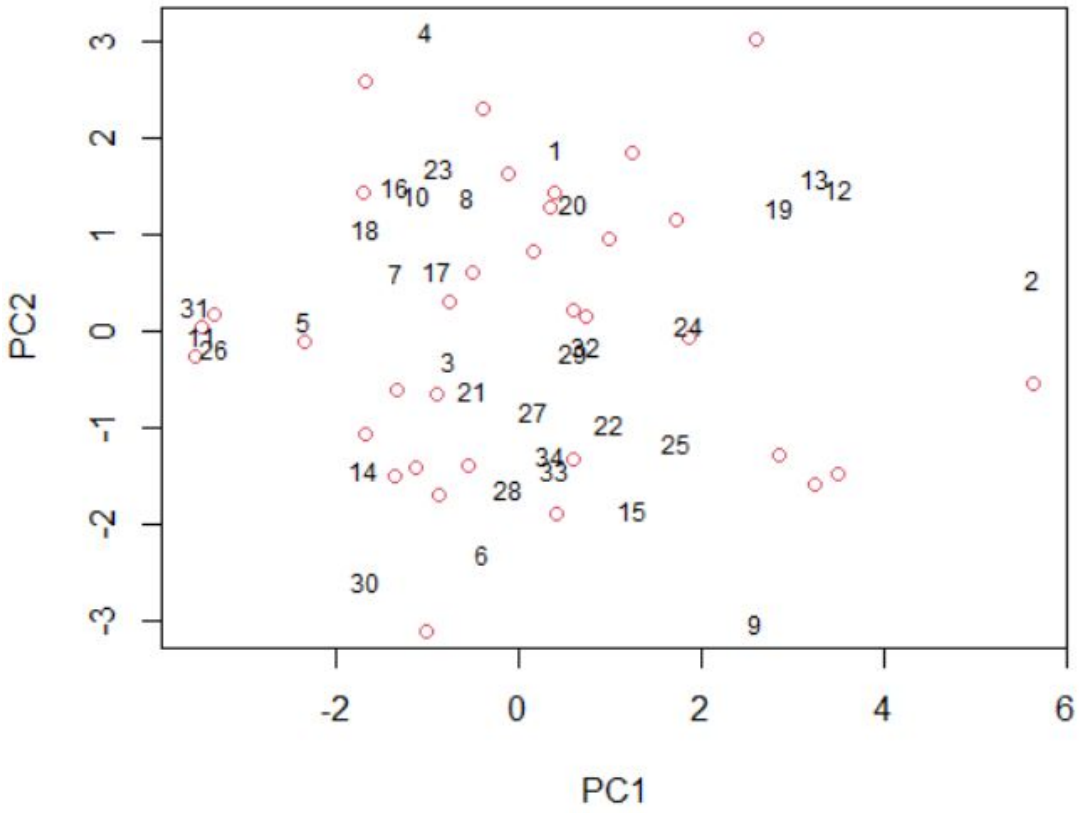
Убеждаемся, что дисперсии, вычесленные разными методами (встроенным и методом главных компонент) близки.

	PC1	PC2
HR.1	0.173	-0.242
SBP.1	-0.014	-0.614
DBP.1	0.206	-0.462
MBP.1	0.027	-0.540
SV.1	-0.464	0.032
CO.1	-0.451	-0.164
SI.1	-0.465	-0.001
CI.1	-0.448	-0.175
TPR.1	0.301	0.016

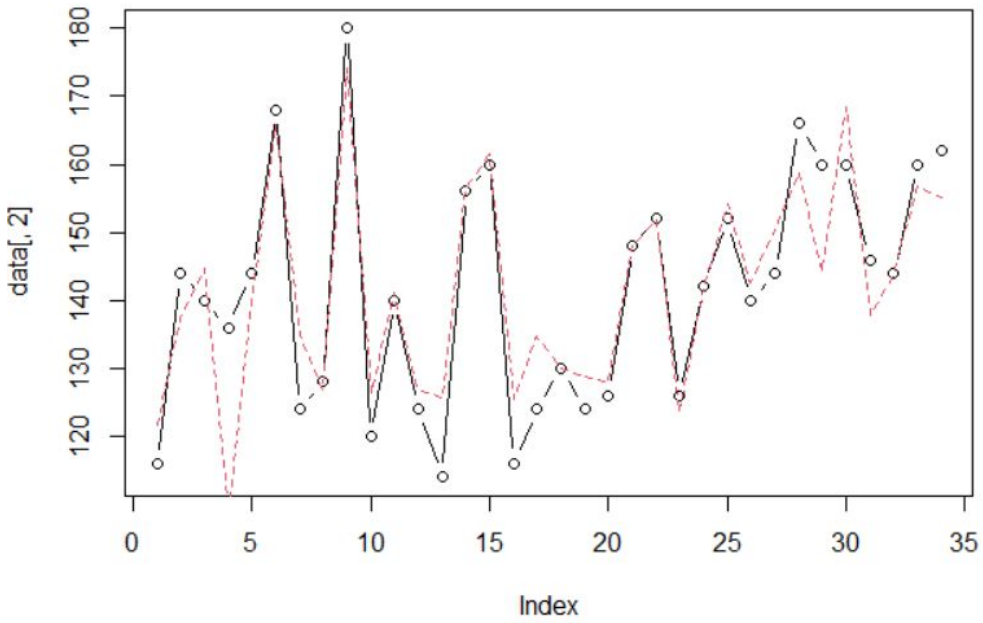
	X1	X2
1	-0.173	-0.242
2	0.014	-0.614
3	-0.206	-0.462
4	-0.027	-0.540
5	0.464	0.032
6	0.451	-0.164
7	0.465	-0.001
8	0.448	-0.175
9	-0.301	0.016

Убеждаемся, что собственные вектора так же близки.

	PC1	PC2
HR.1	0.354	-0.351
SBP.1	-0.030	-0.890
DBP.1	0.423	-0.671
MBP.1	0.056	-0.784
SV.1	-0.953	0.047
CO.1	-0.926	-0.238
SI.1	-0.953	-0.002
CI.1	-0.920	-0.254
TPR.1	0.617	0.023



На этом графике можно хорошо увидеть разницу в работе встроенного и собственноручно описанного метода главных компонент.



По результатам восстановления данных (по двум главным компонентам) можно судить о том, что за исключением нескольких резких расхождений, восстановлению данных по двум компонентам можно доверять.

```
     names    errors
[1,] "HR.1"   "418.863788610954"
[2,] "SBP.1"  "183.890338987723"
[3,] "DBP.1"  "162.099746673394"
[4,] "MBP.1"  "211.669428216228"
[5,] "SV.1"   "261.130297088754"
[6,] "CO.1"   "17.4804255078044"
[7,] "SI.1"   "145.836346978061"
[8,] "CI.1"   "8.87978285917812"
[9,] "TPR.1"  "10559.315877092"
```

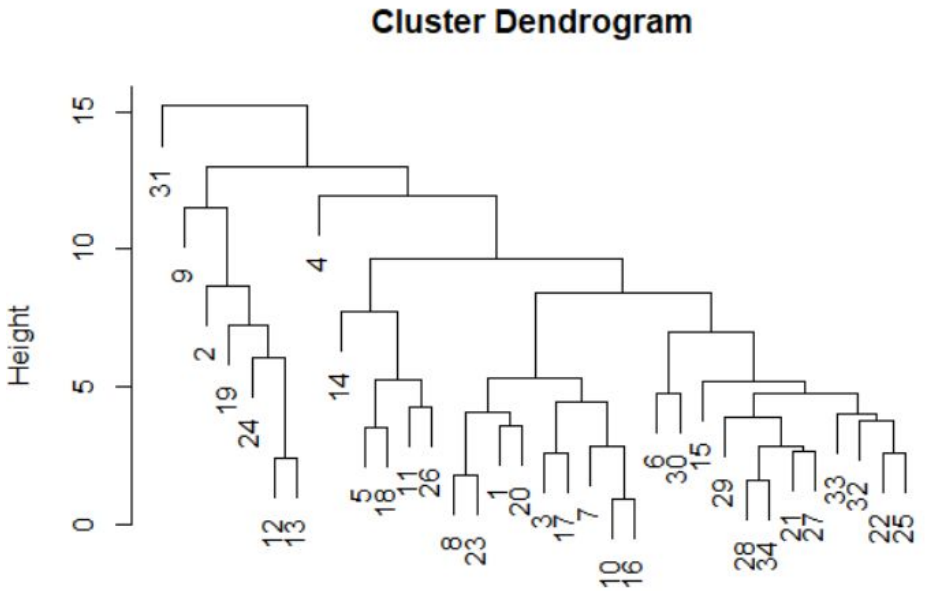
Так, наибольшую ошибку по результатам восстановления можно наблюдать по признакам TPR.1 и HR.1.

Интерпретация факторов

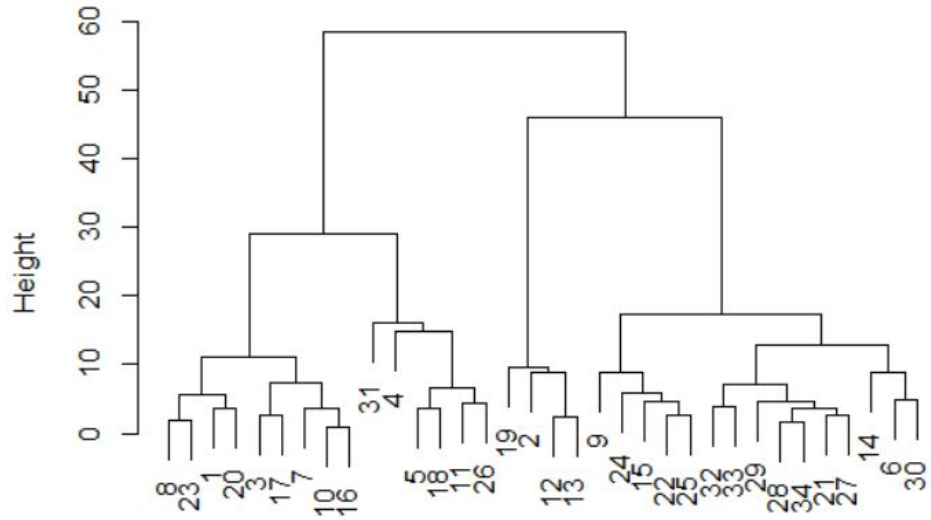
```
[1] "Correlation matrix"
                  craving.to.alcohol.1 tremor.1 sweating.1      1      2
craving.to.alcohol.1                1.000    0.104      0.026 -0.059 -0.505
tremor.1                            0.104    1.000      0.192 -0.068 -0.575
sweating.1                          0.026    0.192      1.000  0.415 -0.108
1                                  -0.059   -0.068      0.415  1.000  0.000
2                                  -0.505   -0.575     -0.108  0.000  1.000
```

Судя по матрице корелляций, можно сделать вывод о том, что второй фактор наибольшим образом линейно пропорционален признаку “tremor.1”, так же, чуть слабее, пропорционален признаку “craving.to.alcohol.1”: чем он больше, тем меньше потливость. А первый фактор наибольшим образом пропорционален признаку “sweating.1”: чем он болше, тем больше потливость и меньше тяга к алгоколю.

Кластерный анализ

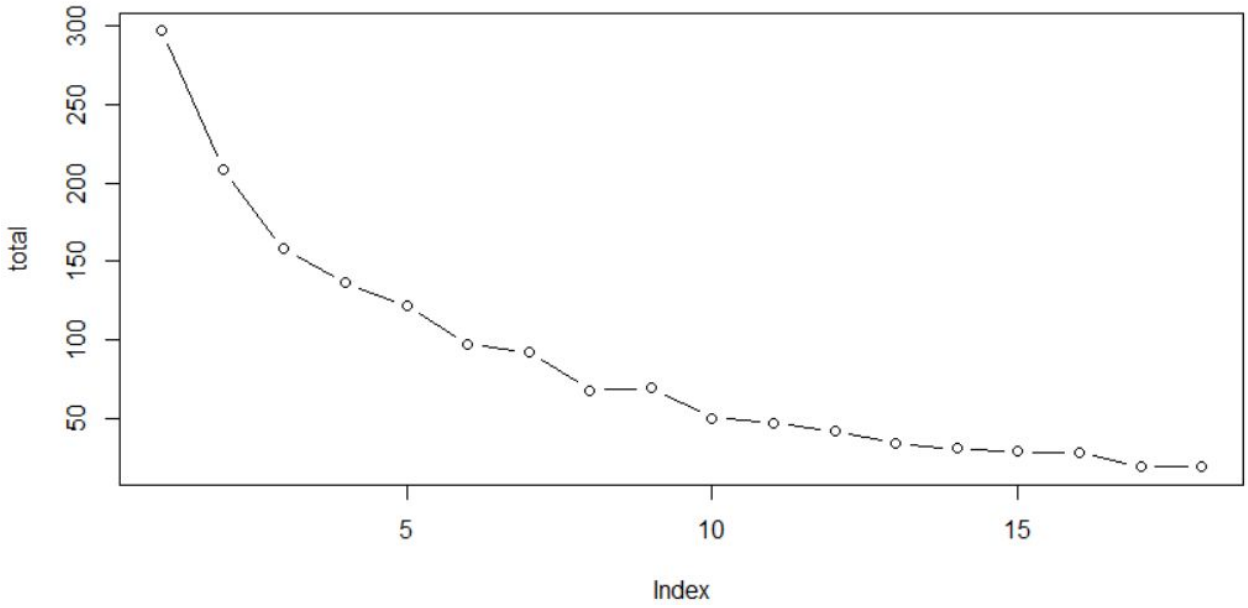


По представленной гистограмме можно судить о том, что наибольшим расстоянием по среднему ото вех остальных и манхеттенской метрике обладают признаки: 31, 9, 4



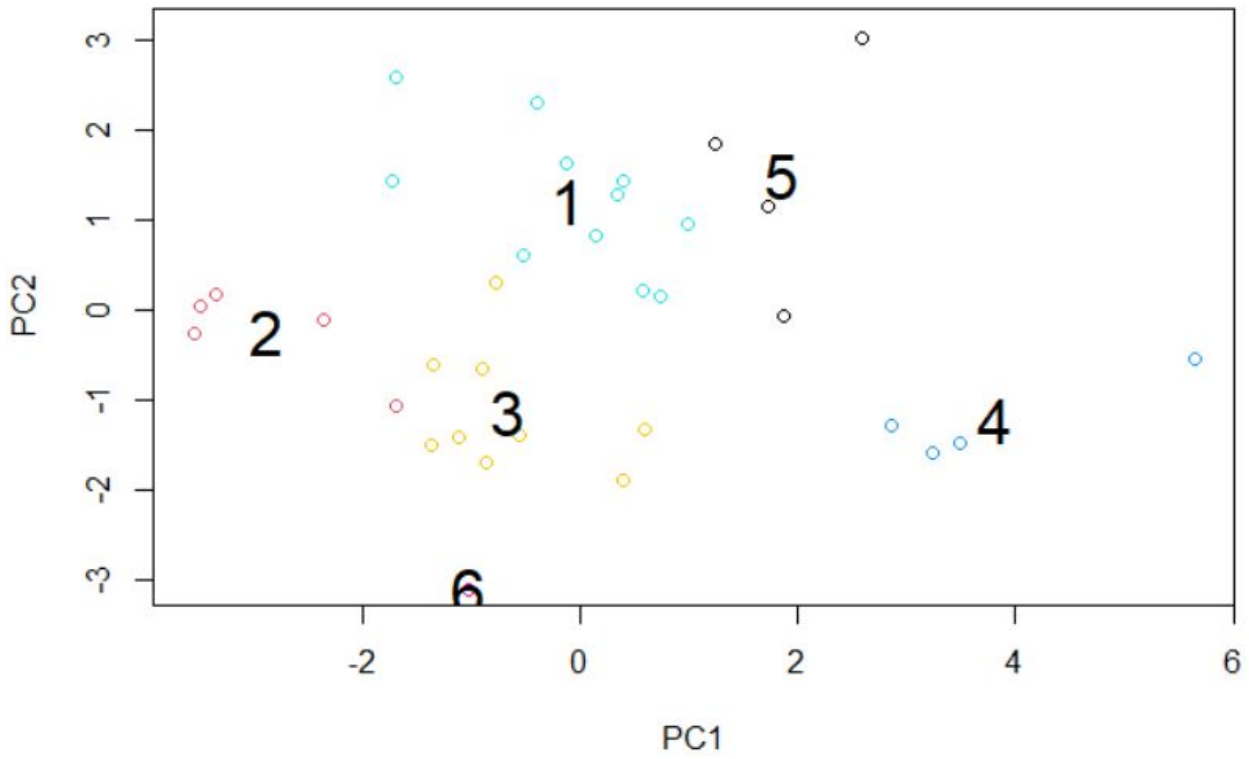
Как видно, с использованием стратегии ward.D, можно добиться соблюдения монотонности и сохранения метрики пространства.

K-means



Выбираем оптимальное количество кластеров по методу локтя. В качестве количества кластеров выберем значение 6.

```
K-means clustering with 6 clusters of sizes 6, 4, 1, 9, 13, 1

Cluster means:
        HR.1      SBP.1      DBP.1      MBP.1        SV.1       CO.1       SI.1       CI.1       TPR.1
1  0.30532275  0.1958934  0.8262989  0.4607294 -0.88718977 -0.9584941 -0.8748140 -0.9165786  1.60879994
2 -0.82923498 -0.8849572 -1.3916614 -0.5471475  1.88291408  1.4571177  1.8838979  1.4872309 -1.00061372
3 -0.31937300  0.2543178  0.3334189  0.4521880 -2.12384328 -2.4737414 -2.1351392 -2.5477351 -2.36178028
4  0.34206956 -1.0050517 -0.6523413 -0.3450027 -0.40901708 -0.3606912 -0.4609294 -0.3728598 -0.01087483
5  0.01134828  0.8835032  0.4168293  0.5270889  0.28249419  0.4925622  0.3294968  0.4966501 -0.31818499
6 -1.42177727 -0.3299258  0.7277229 -4.7751058 -0.07594507 -0.7608531 -0.2866622 -1.0024290  0.94571377

Clustering vector:
 [1] 4 2 4 6 1 5 4 4 5 4 1 2 2 1 5 4 4 1 2 4 5 5 4 5 5 1 5 5 5 1 3 5 5 5

Within cluster sum of squares by cluster:
[1] 23.23723 11.42165  0.00000 18.39855 35.03623  0.00000
 (between_SS / total_SS =  70.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```



Результаты кластеризации. Как можно увидеть, кластеры были поделены аккуратно и относительно равномерно.