Кластерный анализ

Количество кластеров?

Валидация кластеров

Количество кластеров? Графически

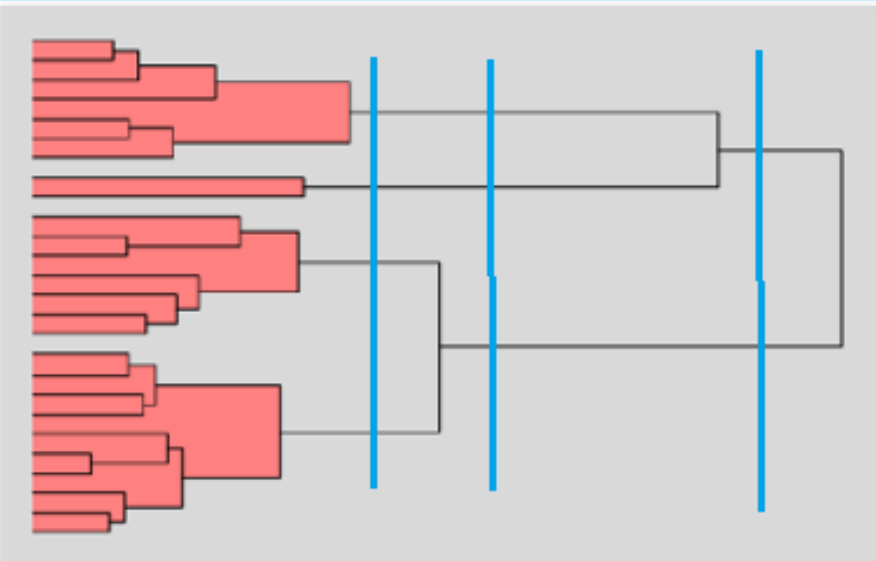
- Главное правило по выбору числа кластеров: выбрать столько, сколько можно интерпретировать содержательно.

Шаг 1.

- Иерархический метод, **дендрограмма**
- Метод k-means, **Elbow method** (“метод согнутого колена”, он же “метод каменистой осыпи”).

Графически - дендрограмма

- Иерархический метод, дендрограмма, выбрать число кластеров

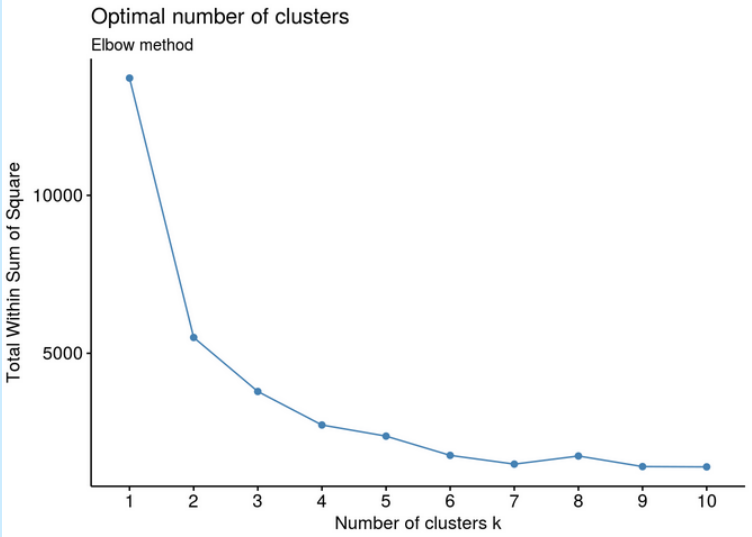


2, 3, 4?

Графически - Elbow method

Строится график, где по оси абсцисс отмечено число кластеров k,
а по оси ординат – значения функции W(k),
которая определяет внутригрупповой разброс
в зависимости от числа кластеров. Кластеры находятся методом k-means.

2, 3, 4 ?



В качестве W(k) можно выбрать F_2 или F_3 (возможны и другие критерии)

l - номер кластера;
d - расстояние

1. Сумма внутриклассовых расстояний между объектами:

$$F_2 = \sum_l \sum_{ij} d^2_{ij}$$

2. Суммарная внутриклассовая дисперсия:

$$F_3 = \sum \sum \sigma_{lj}^2,$$

где σ_{lj}^2 - дисперсия j-ой переменной в кластере S_l

- Шаг 2.
- Метод k-means, для отобранных k.
- Окончательно: выбрать столько, сколько можно интерпретировать содержательно.

Валидация кластеров

- Валидация – это проверка продукта, процесса или системы на соответствие требованиям клиента.
- **Например, в R:**
- Silhouette () вычисляет или извлекает информацию о силуэте
- cluster.stats () вычисляет несколько статистических данных валидности кластера из кластеризации и матрицы несходства (fpc). (Несходство можно определить как расстояние между двумя образцами по некоторому критерию, другими словами, насколько эти образцы различны. Например, евклидово расстояние между двумя точками является мерой их несходства).
- clValid () вычисляет меры проверки для заданного набора алгоритмов кластеризации и количества кластеров (clValid)
- clustIndex () вычисляет значения нескольких индексов кластеризации, которые можно независимо использовать для определения количества кластеров, существующих в наборе данных (cclust)
- NbClust () предоставляет 30 индексов для проверки кластера и определение количества кластеров (NbClust)

Метод силуэта (Silhouette)

- Значение силуэта для каждого объекта является мерой того, насколько он является «хорошим» для своего кластера. То есть насколько объект похож на его собственный кластер по сравнению с другими кластерами.
- Значение силуэта для *i*-ого объекта определяется по формуле:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ если } |C_i| > 1 \quad \text{и} \quad s(i) = 0, \text{ если } |C_i| = 1$$

- здесь $a(i)$ – среднее расстояние от *i*-того объекта до объектов своего кластера,
- $b(i)$ – среднее расстояние от *i*-того объекта до объектов ближайшего соседнего кластера.

Метод силуэта (Silhouette)

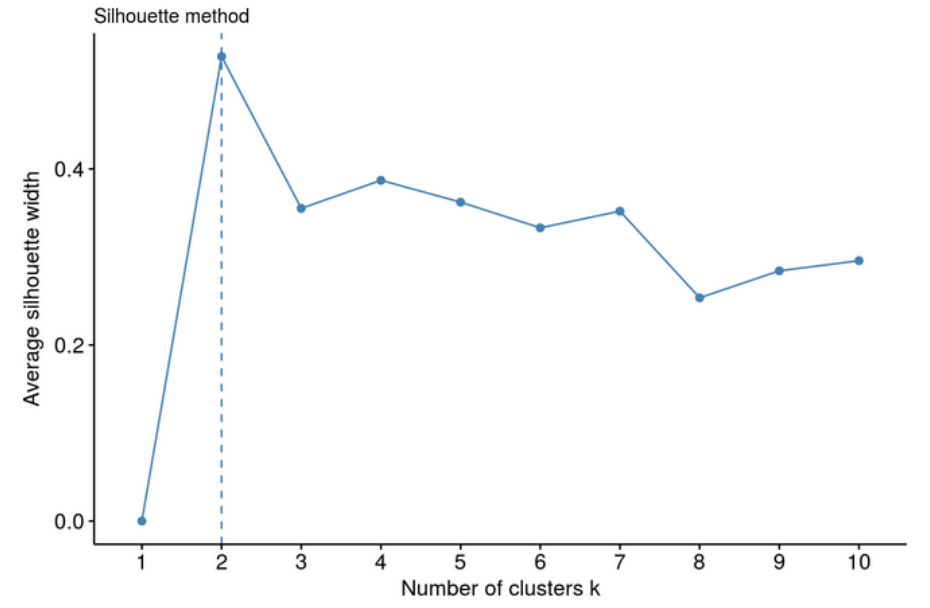
- Значение $s(i)$ изменяется в пределах от -1 до +1. Оценка равна 0 для кластеров с размером = 1.
- Если силуэт равен 1, то кластеризация успешна.
- Если это 0, то, скорее всего, кластеры выбраны неверно.
- Если же силуэт меньше нуля, то скорее всего объект должен быть в другом кластере. В этом случае нужно что-то поменять, например, количество кластеров, измерение расстояний.
- Кауфман ввел термин **силуэтный коэффициент** - максимальное значение усредненное значение силуэта по всем кластерам данного набора данных.

$$SC = \max_k \tilde{s}(k)$$

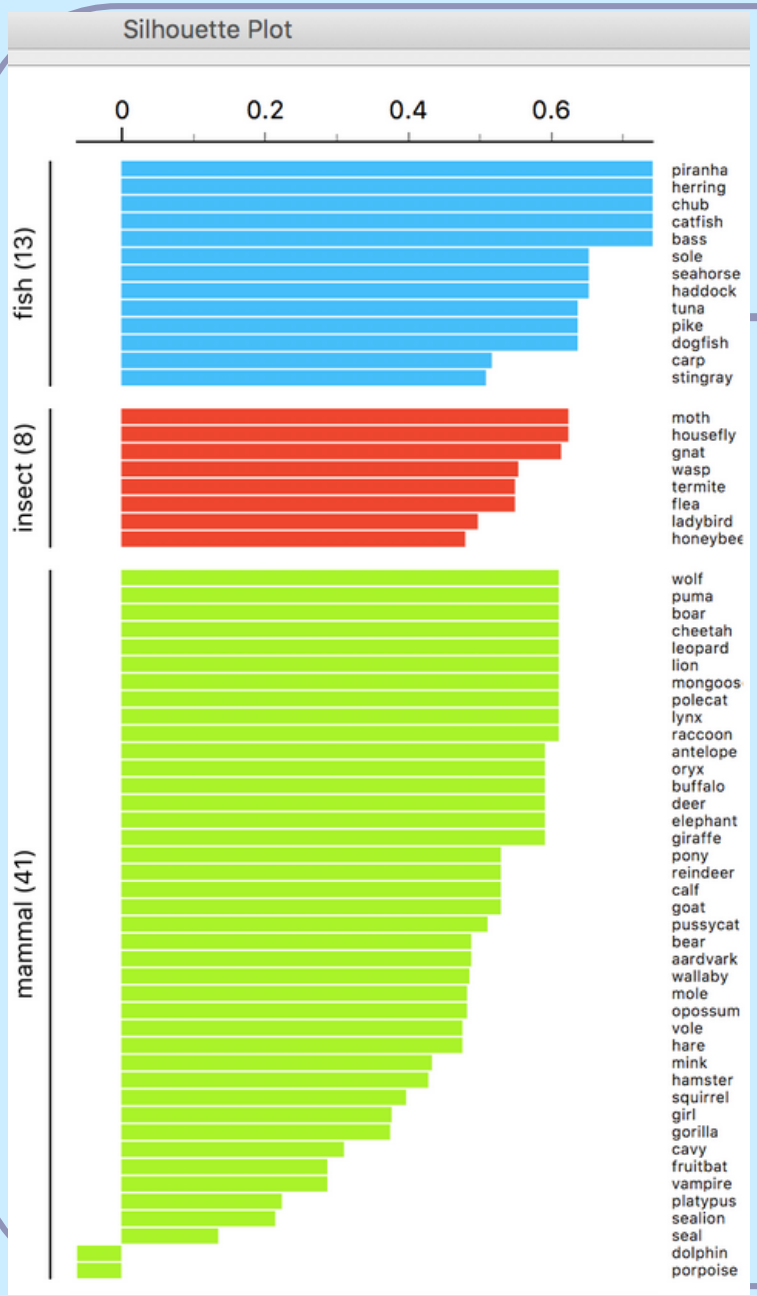
- Значение силуэта можно рассматривать, как степень пересечения кластеров **друг с другом**, то есть -1: перекрываются, +1: кластеры совершенно разделимы.

Метод силуэта (Silhouette)

- Можно сравнить полученные значения в зависимости от количества кластеров k: а)усредненное значение силуэта для всего набора данных, б) силуэтный коэффициент.



Метод силуэта (Silhouette)



Результаты кластеризации –
рыбы,
насекомые,
млекопитающие.

Пример

- Файл **European.xlsx**
- Страны

{"Belgium","Denmark","France","W_Germany","Ireland","Italy","Luxembourg",
"Netherlands","United_Kingdom","Austria","Finland","Greece","Norway",
"Portugal","Spain","Sweden","Switzerland","Turkey","Bulgaria","Czechoslovakia",
"E_Germany","Hungary","Poland","Rumania","USSR","Yugoslavia"}

- Доля занятых в секторе:

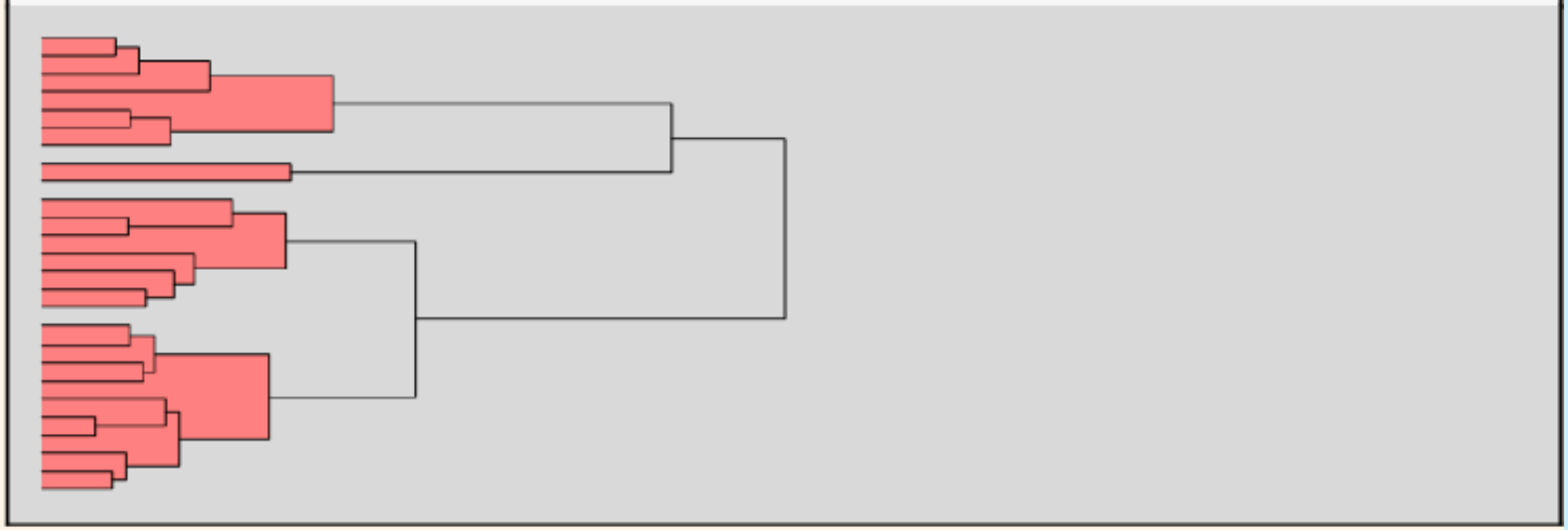
Agr - с/х, **Min** - горно-добывающей пром-ти, **Man** – промышленности,
PS – энергентика, **Con** – строительстве, **SI -** сфера услуг , **Fin** - финансовый сектор, **SPS** - соц.службы, **TC** - транспорт и связь.

КЛАСТЕРНЫЙ АНАЛИЗ

Иерархический кластерный анализ

```
Clear[distanceMatrix]
Needs["HierarchicalClustering`"]
(distanceMatrix = DistanceMatrix[stdata, DistanceFunction → EuclideanDistance]) //
  MatrixForm;
Clear[clust]
clust = DirectAgglomerate[distanceMatrix, countries, Linkage → "Ward"];
```

```
plot = DendrogramPlot[clust, Orientation → Right, HighlightLevel → 4,
  HighlightStyle → Pink]
```



Получили 4 кластера, но сразу можно увидеть, что они не очень однородные по количеству стран - есть один кластер, который довольно маленький (2 страны).

Метод к - средних 4 или 3 кластера (хочется убрать кластер, который состоит из 2 стран).

```
clustKmeans = FindClusters[stdata, 4, Method -> "KMeans"];
clustK1 = clustKmeans[[1]];
clustK2 = clustKmeans[[2]];
clustK3 = clustKmeans[[3]];
clustK4 = clustKmeans[[4]];
Map[Length, {clustK1, clustK2, clustK3, clustK4}]
```

```
{14, 5, 1, 6}
```

Если разбивать на 4 кластера, то группы получаются достаточно неоднородные по количеству, попробуем 3.

```
clustKmeans2 = FindClusters[stdata, 3, Method -> "KMeans"];
clustK12 = clustKmeans2[[1]];
clustK22 = clustKmeans2[[2]];
clustK32 = clustKmeans2[[3]];
Map[Length, {clustK12, clustK22, clustK32}]
```

```
{17, 2, 7}
```

Сначала 3 кластера:

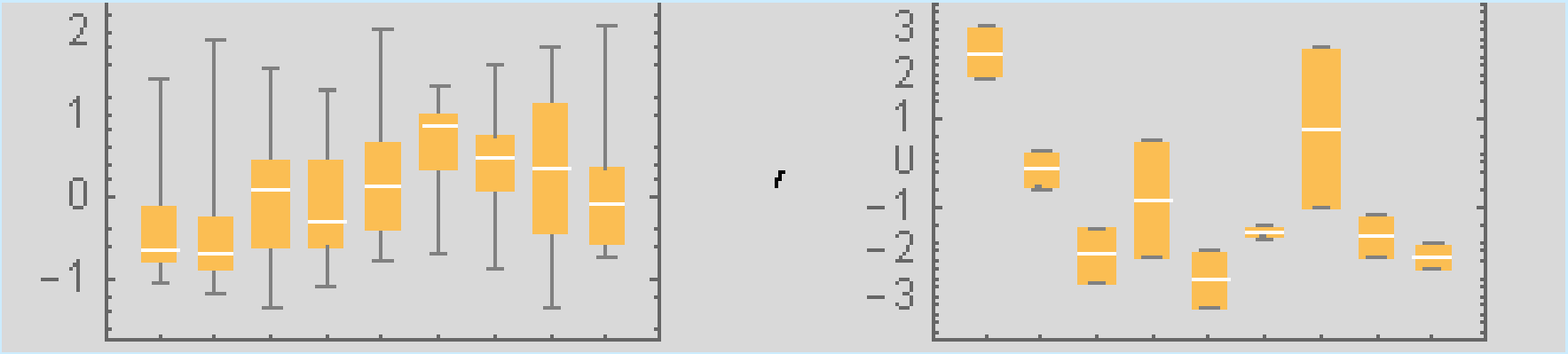
```
pos1 = Flatten@Table[Position[stdata, i], {i, clustK12}];
pos2 = Flatten@Table[Position[stdata, i], {i, clustK22}];
pos3 = Flatten@Table[Position[stdata, i], {i, clustK32}];
country1 = Table[countries[i], {i, pos1}]
country2 = Table[countries[i], {i, pos2}]
country3 = Table[countries[i], {i, pos3}]
```

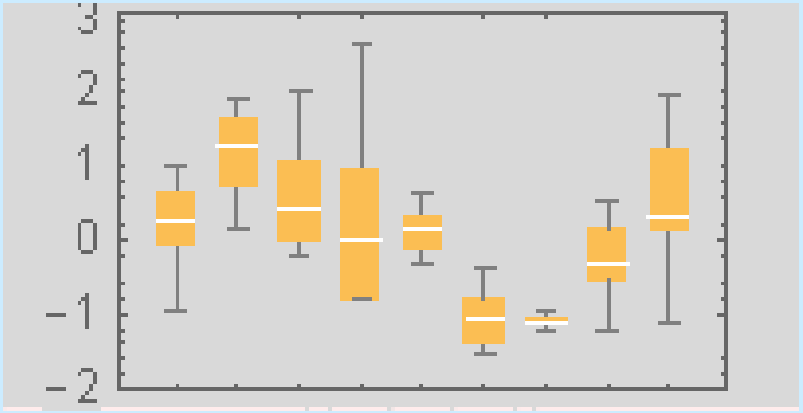
```
{Belgium, Denmark, France, W_Germany, Ireland,
 Italy, Luxembourg, Netherlands, United_Kingdom, Austria,
 Finland, Greece, Norway, Portugal, Spain, Sweden, Switzerland}
```

```
{Turkey, Yugoslavia}
```

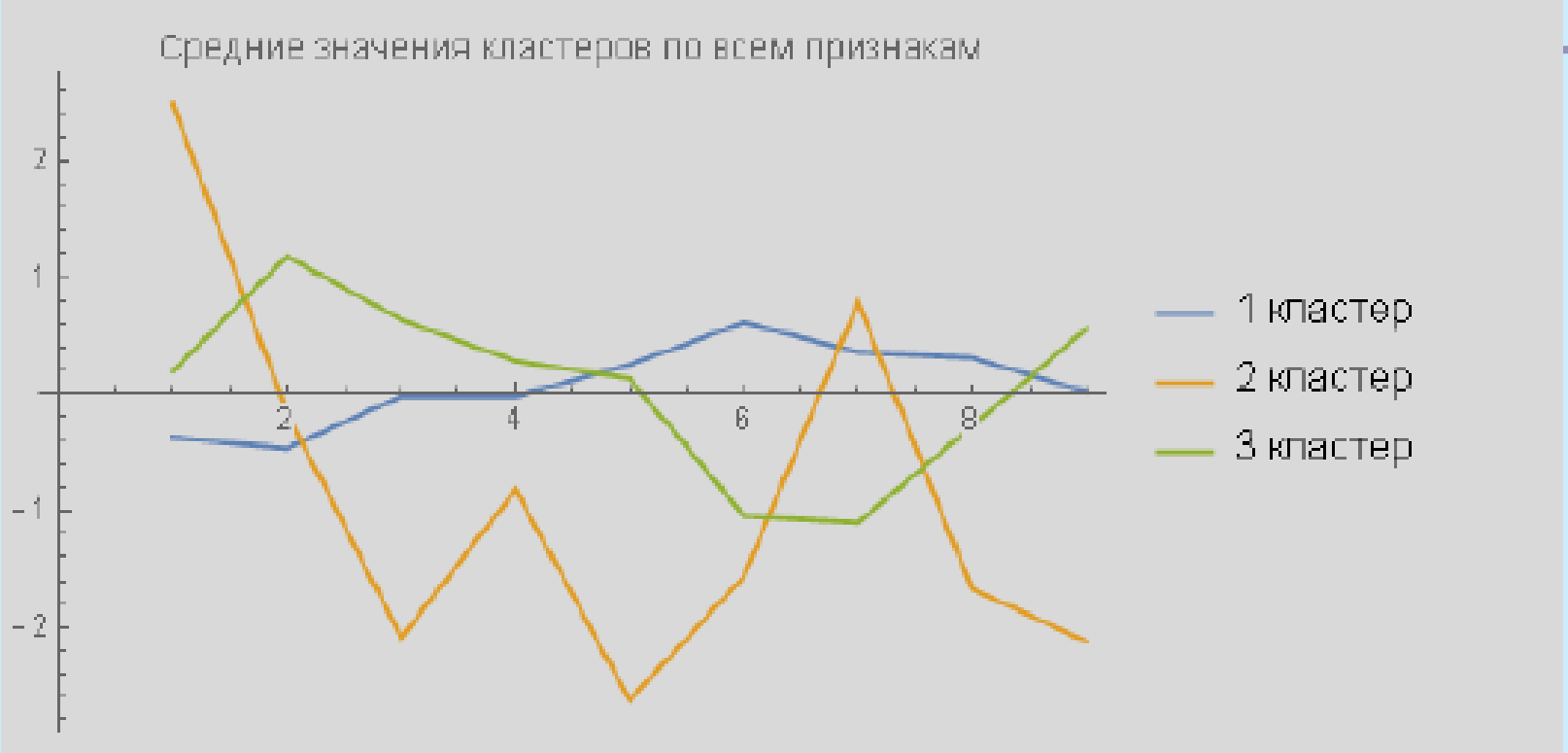
```
{Bulgaria, Czechoslovakia, E_Germany, Hungary, Poland, Rumania, USSR}
```

Распределение признаков Boxplot





```
ListLinePlot[Mean /@ Table[clustKmeans2[[i]], {i, 1, 3}],
 PlotLegends → {"1 кластер", "2 кластер", "3 кластер"},
 PlotLabel → "Средние значения кластеров по всем признакам"]
```



- Сравнение средних показателей
- Выводы: описание кластеров