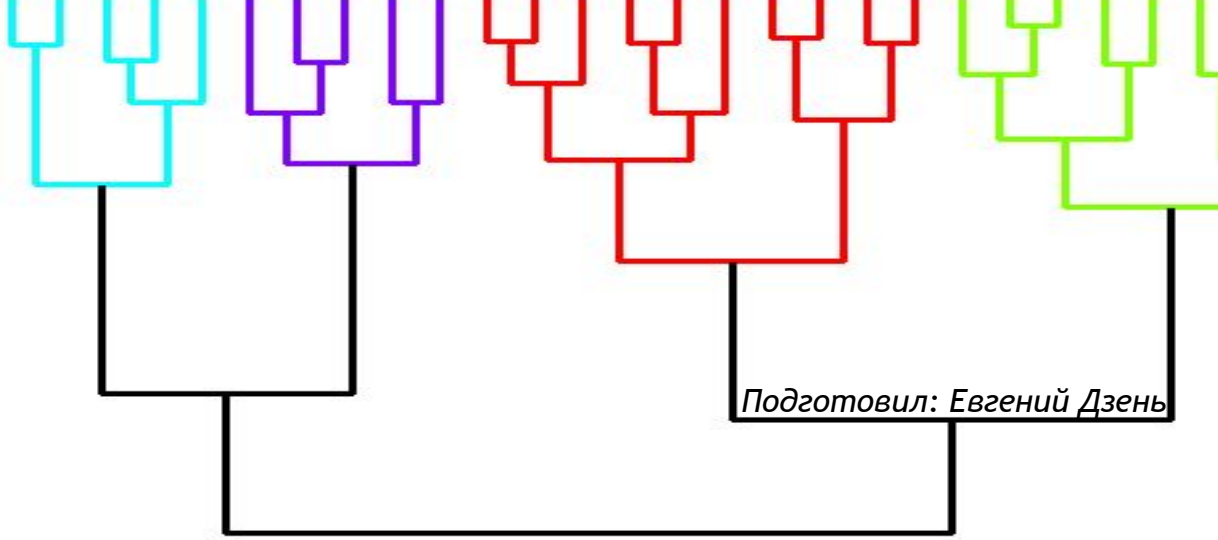


Подготовил: Евгений Дзень

Кластерный анализ

1. Суть кластерного анализа
2. История возникновения метода
3. Рассмотрение типичной задачи

   (с использованием STATISTICA 8.0)
4. Методы кластерного анализа и его специфика
5. Меры расстояния
6. Алгоритмы объединения в кластеры
7. Рассмотрение примера из сферы бизнеса

1. **Суть кластерного анализа**
2. История возникновения метода
3. Рассмотрение типичной задачи

   (с использованием STATISTICA 8.0)
4. Методы кластерного анализа и его специфика
5. Меры расстояния
6. Алгоритмы объединения в кластеры
7. Рассмотрение примера из сферы бизнеса

Древняя китайская классификация животных

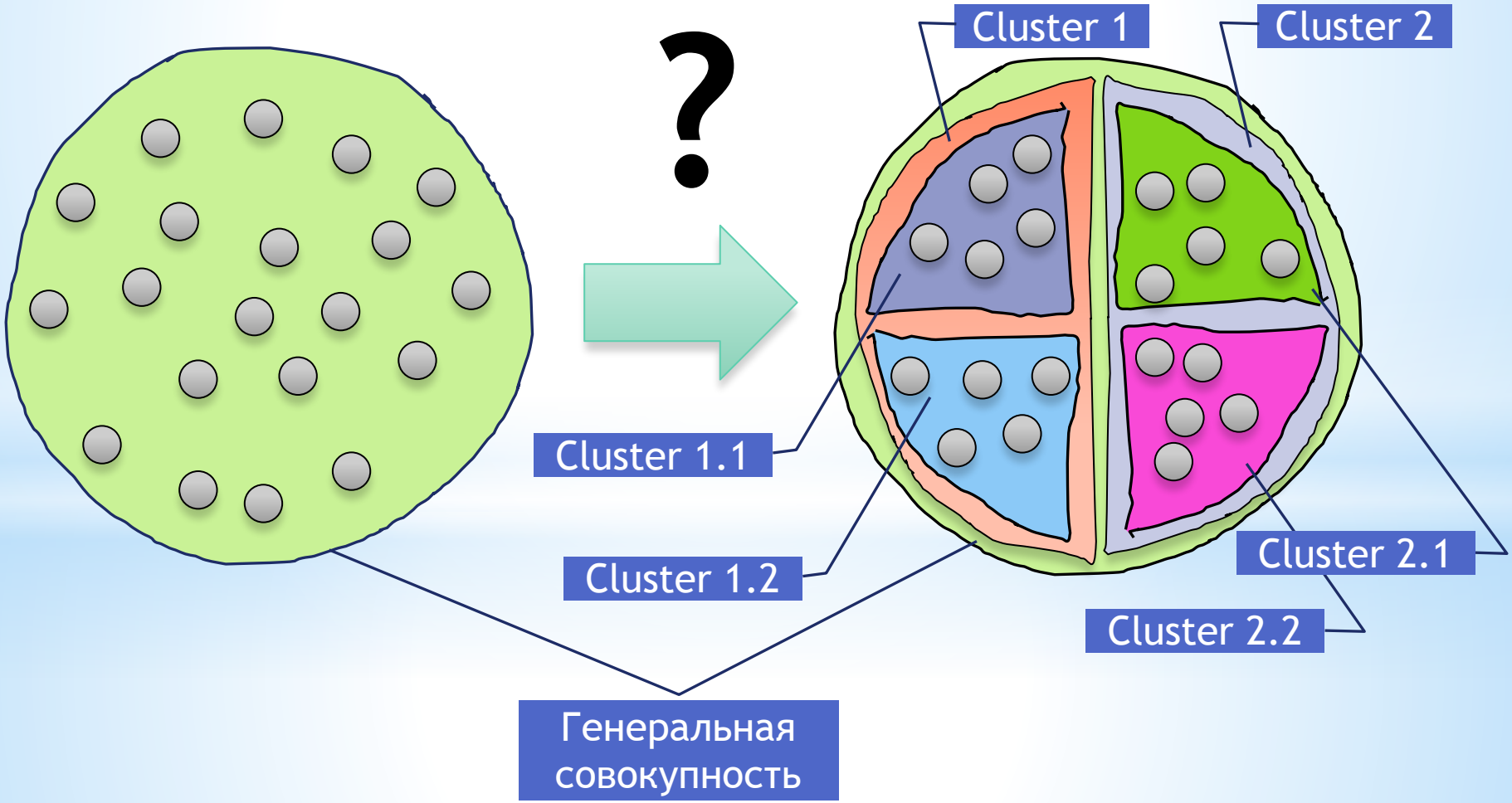
Животные подразделяются на:

а) принадлежащих императору;
б) набальзамированных;
в) дрессированных;
г) молочных поросят;
д) сирен;
е) сказочных;
ж) бродячих собак;
з) включённых в данную классификацию;
и) дрожащих, как сумасшедшие;
к) неисчислимых;
л) нарисованных самой лучшей верблюжьей кисточкой;
м) других;
н) тех, которые только что разбили цветочную вазу и
о) тех, которые издалека напоминают мух.

(Хорхе Луис Борхес, *Другие исследования:* 1937—1952).

Задача разбиения на классы…

Как определить, к какому классу отнести тот или иной элемент генеральной совокупности, характеризующийся множественными параметрами?

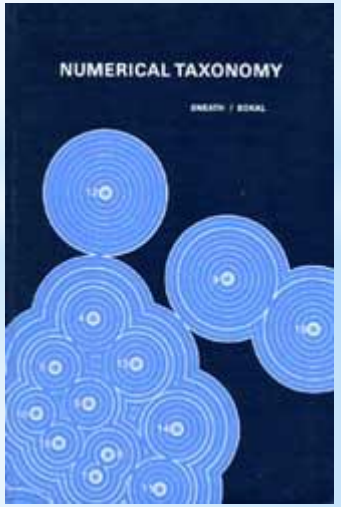


1. Суть кластерного анализа
2. **История возникновения метода**
3. Рассмотрение типичной задачи

   (с использованием STATISTICA 8.0)
4. Методы кластерного анализа и его специфика
5. Меры расстояния
6. Алгоритмы объединения в кластеры
7. Рассмотрение примера из сферы бизнеса

Истоки…

- Первые работы, описывающие методы кластерного анализа относятся к концу 30-х годов.
- Считается, что термин «кластерный анализ» первым в употребление ввёл американский психолог из университета Беркли ***Роберт Трайон (Robert C. Tryon)*** в 1939.
- Однако активный интерес к данной теме пришёлся на период 60-80 гг.
- Импульсом для разработки многих кластерных методов послужила книга «**Начала численной таксономии**», опубликованная в 1963 г. двумя биологами — **Робертом Сокэлом** и **Петером Снитом** **(Sneath, Sokal)**.



1. Суть кластерного анализа
2. История возникновения метода
3. **Рассмотрение типичной задачи**

   (с использованием STATISTICA 8.0)
4. Методы кластерного анализа и его специфика
5. Меры расстояния
6. Алгоритмы объединения в кластеры
7. Рассмотрение примера из сферы бизнеса

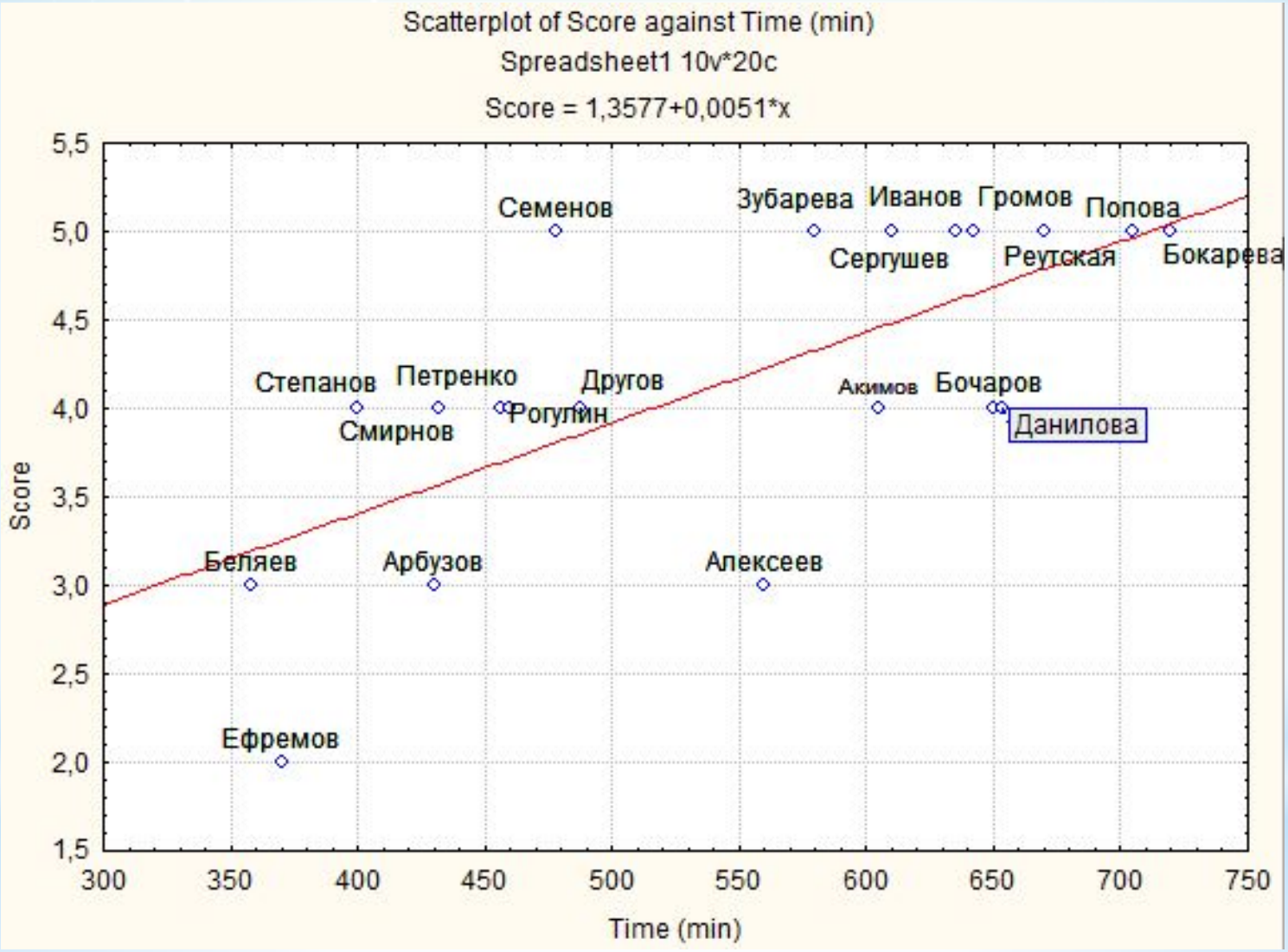
Входные данные

- В исходной таблице мы имеем данные по группе студентов за истекший семестр
- Проведя регрессионный анализ, мы выяснили, что между двумя параметрами (Time, Score) имеется устойчивая положительная линейная зависимость (коэффициент корреляции Пирсона = 0,68 при α=0,05)
- Взглянем на наши данные построив диаграмму рассеяния…

Data: Spreadsheet1* (10v by 20c)

	1 Time (min)	2 Score
Акимов	605	4
Алексеев	560	3
Арбузов	430	3
Беляев	358	3
Бокарева	720	5
Бочаров	650	4
Громов	642	5
Данилова	654	4
Другов	488	4
Ефремов	370	2
Зубарева	580	5
Иванов	635	5
Петренко	456	4
Попова	705	5
Реутская	670	5
Рогулин	460	4
Семенов	478	5
Сергушев	610	5
Смирнов	432	4
Степанов	400	4

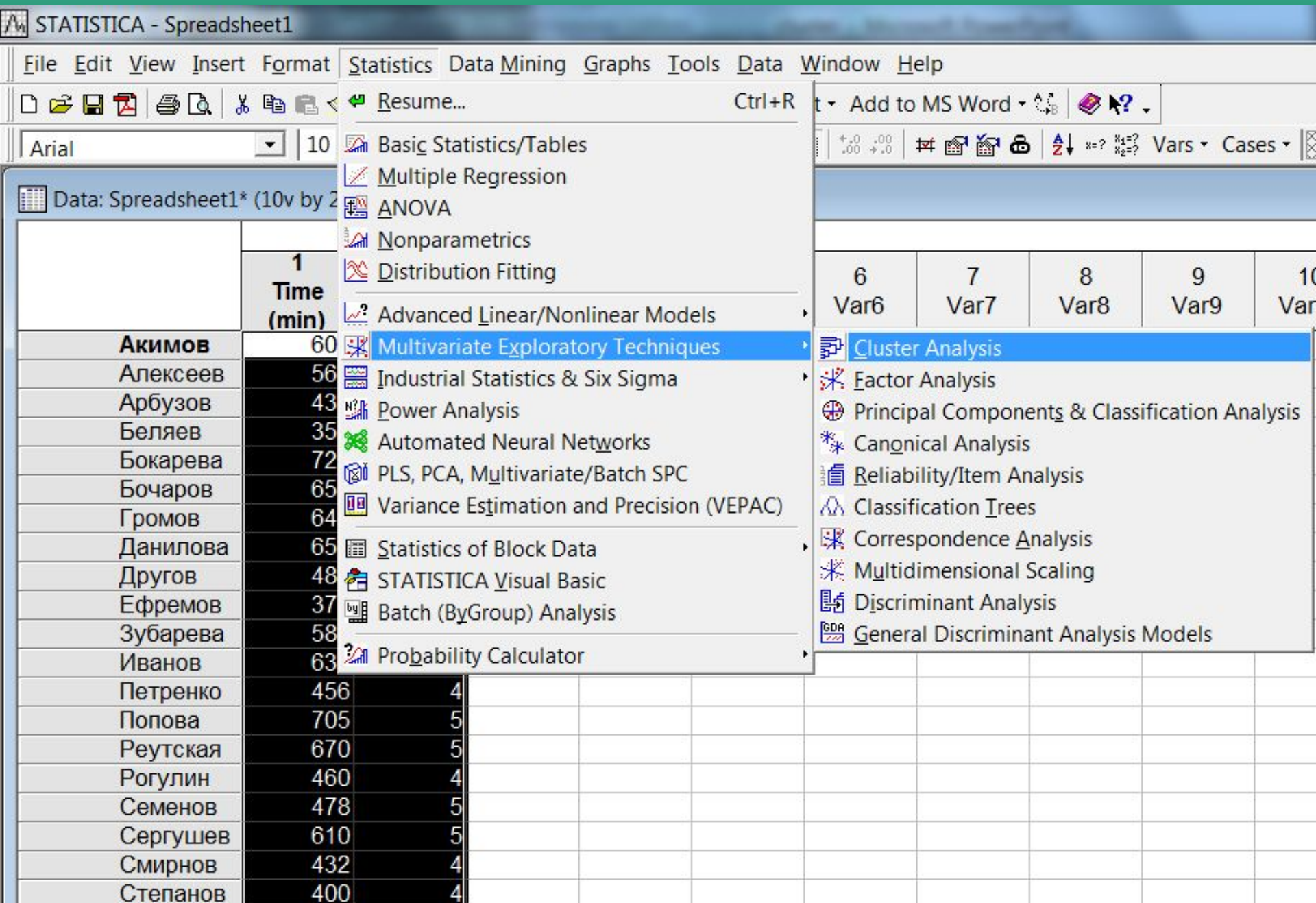
Диаграмма рассеяния объектов наблюдений



Как можно охарактеризовать такую неоднородность?

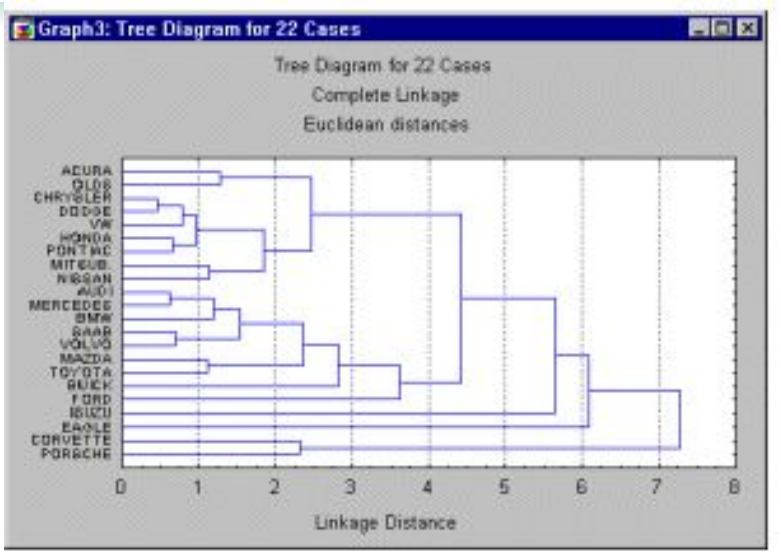
Какие группы объектов можно выделить?

Вызов инструмента **«Cluster Analysis»**



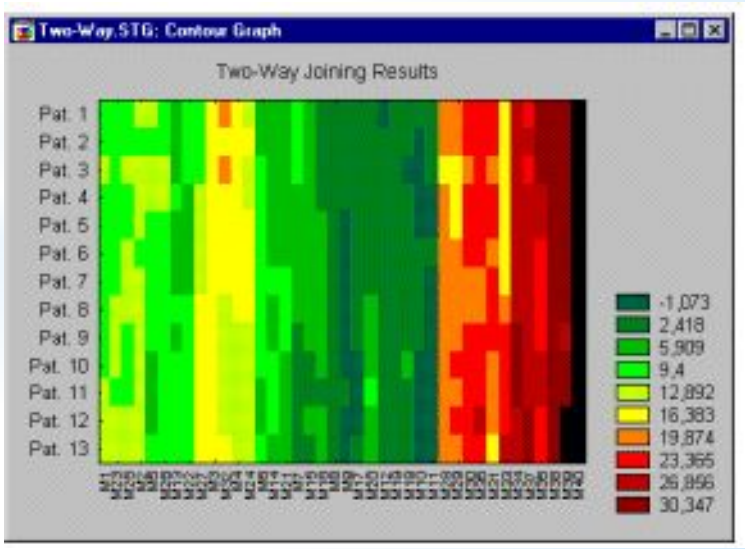
Выбор метода кластеризации

- Древовидная кластеризация

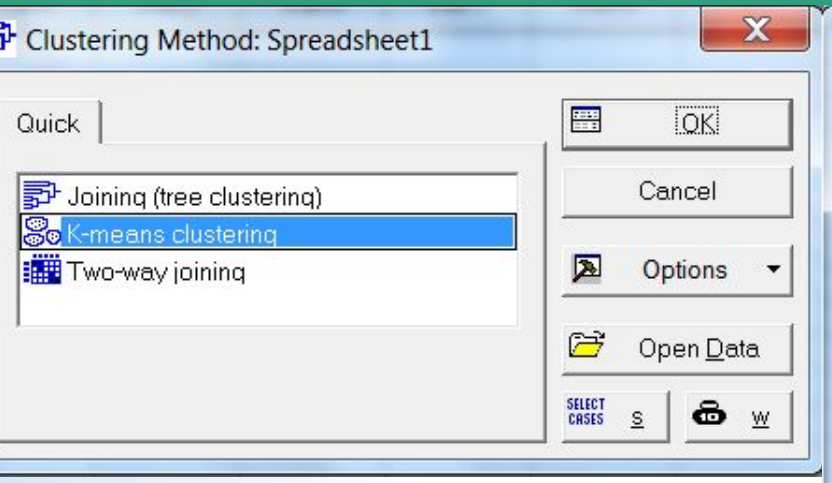


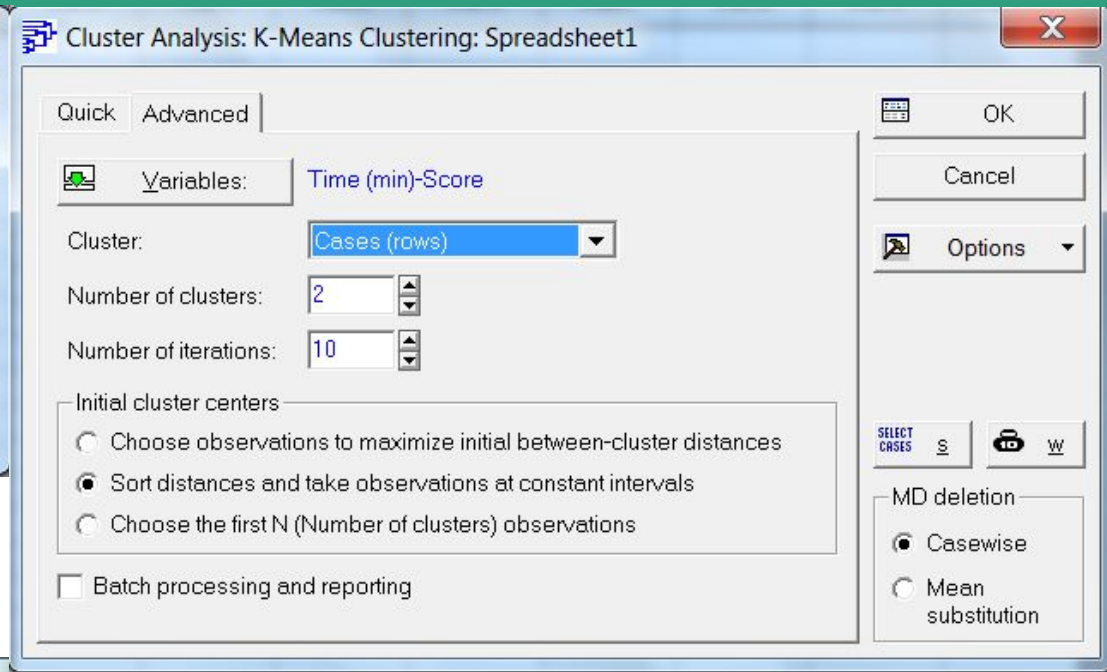
- **Кластеризация по методу K-средних**

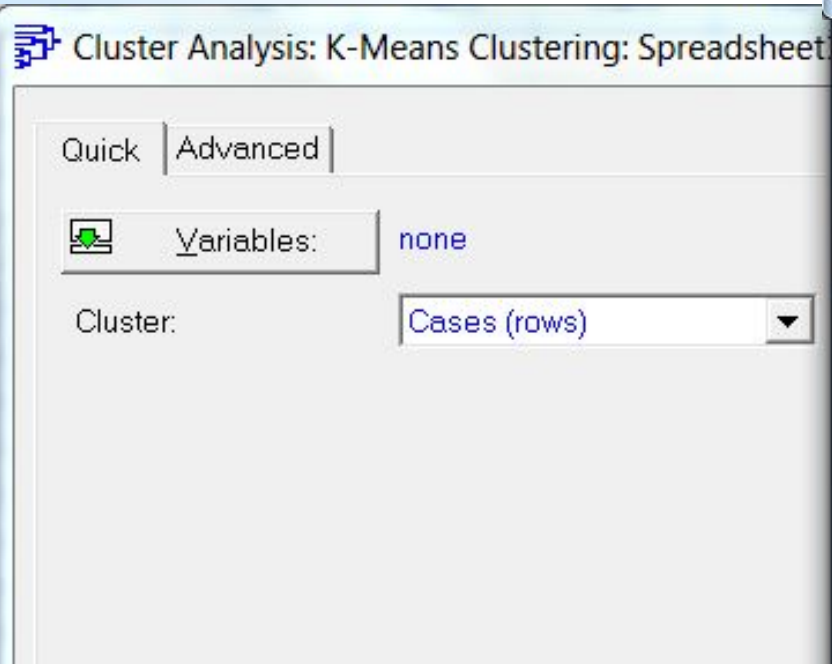
- Двувходовое объединение

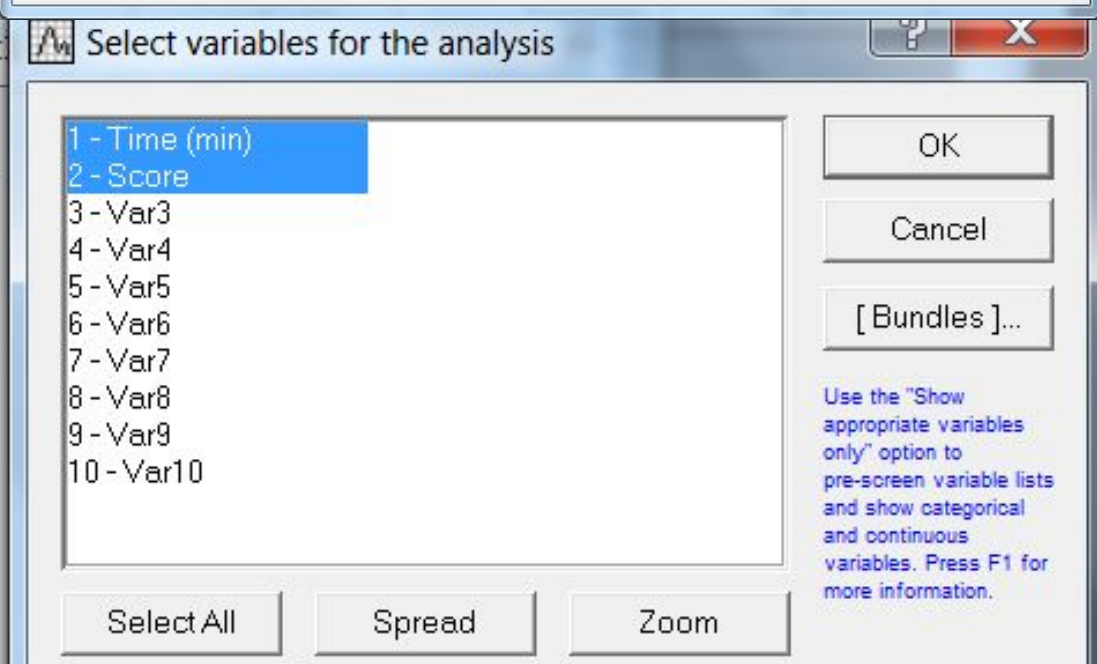


Задание параметров кластеризации









А сколько кластеров?!..

Не существует единственно правильной априорной разбивки на кластеры. Поэтому нужно пробовать разные варианты разбивки.

Выделяют два критерия «хорошей» разбивки на кластеры:

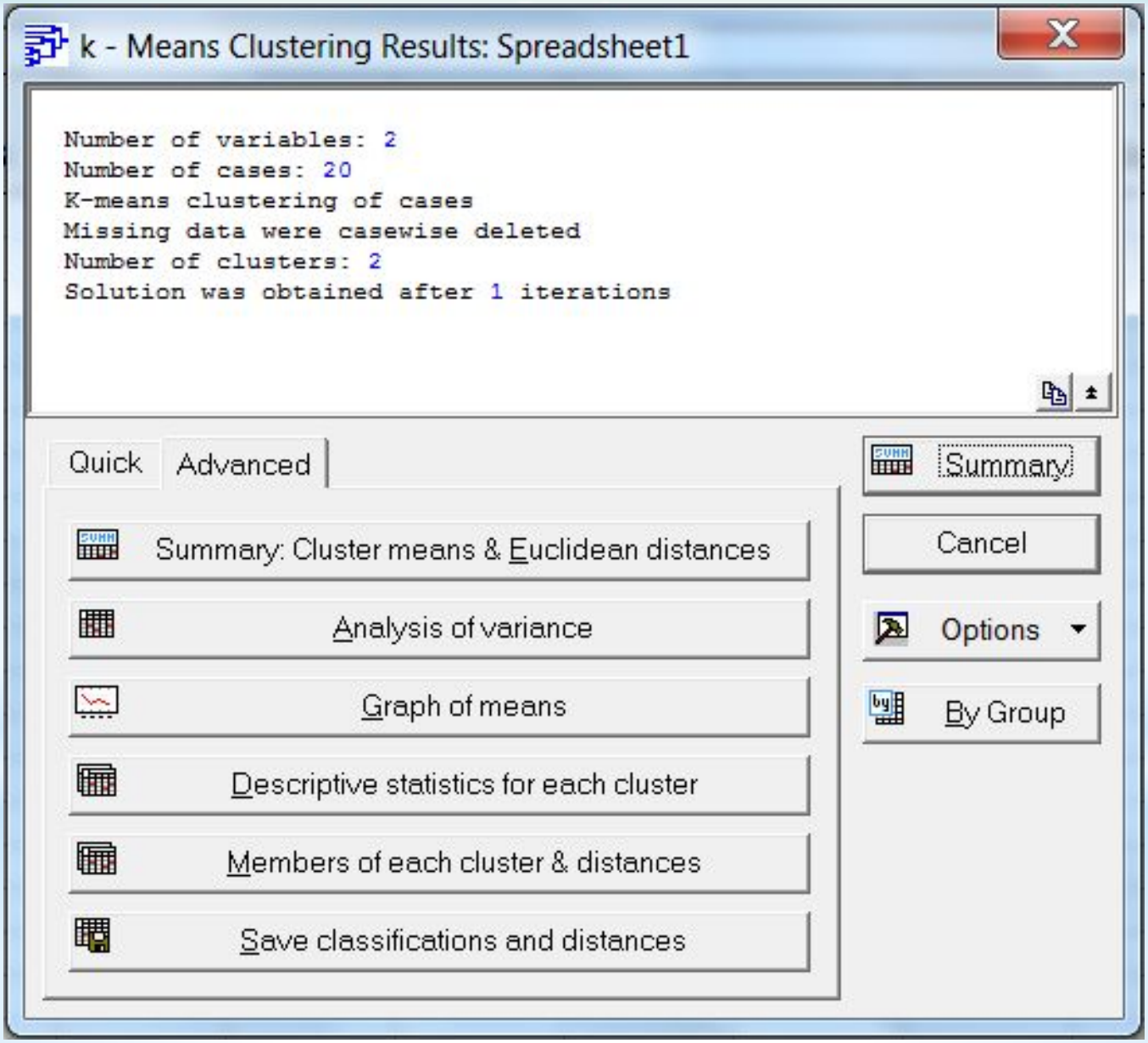
ПЕРВЫЙ — формальный —

связан с тем, что объекты одной группы заметно отличаются от объектов другой группы по всем включенным в анализ переменным;

ВТОРОЙ — содержательный —

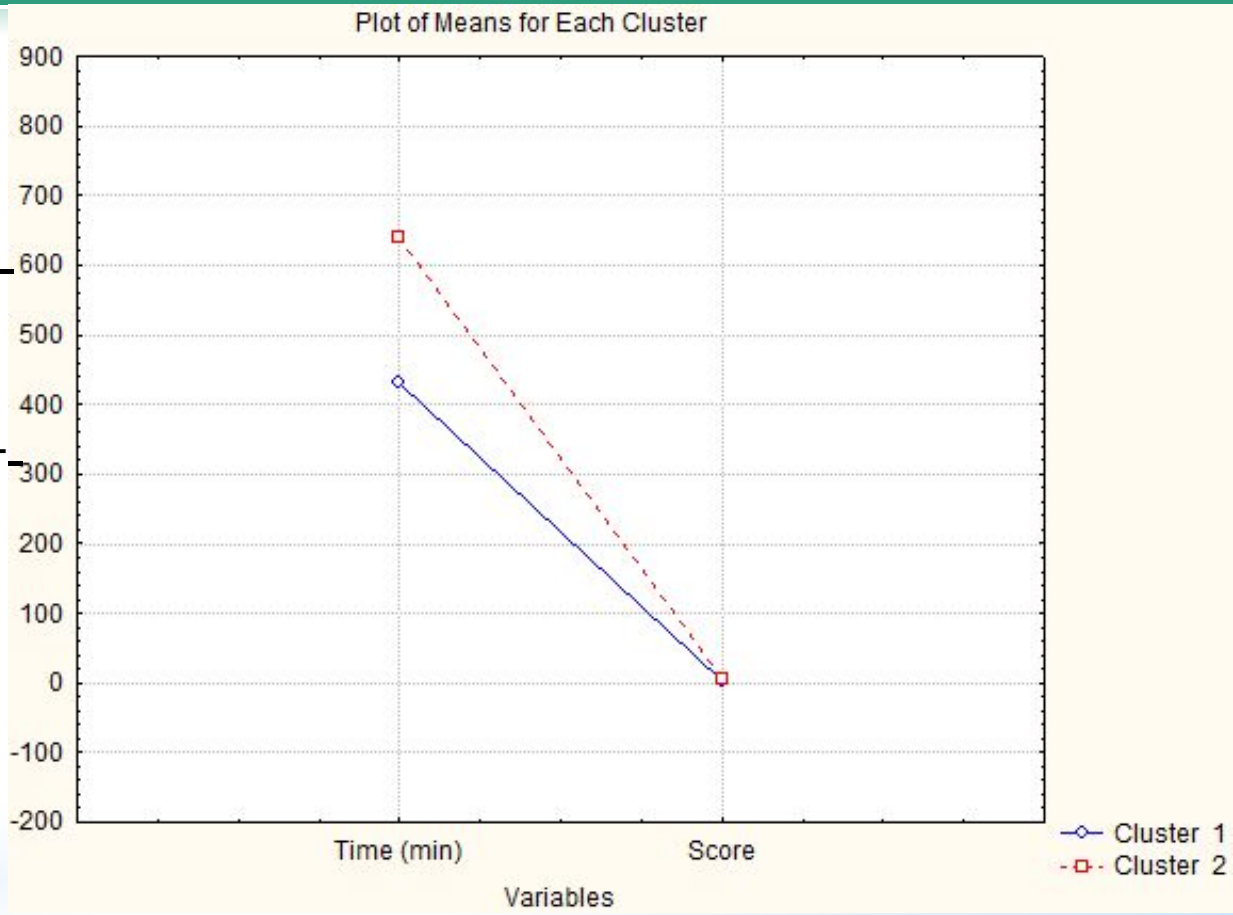
определяется возможностью разумной интерпретации каждого кластера.

Вывод результатов



И что же вышло? =(

График показывает, что кластеры заметно отличаются по переменной «время» и практически не отличаются по переменной «оценка». Таким образом, вторая переменная является как бы лишней, не добавляя никакой информации. **Почему так происходит?**



Обратим внимание на то, что для измерения переменной «время» используются трехзначные числа, а для переменной «оценка» — одноразрядные.

Решение данной проблемы – стандартизация данных!

Стандартизация данных

Как сделать переменные равноправными в образовании кластеров?

1. **Вычислим среднее арифметическое и стандартное отклонение каждой из переменных**

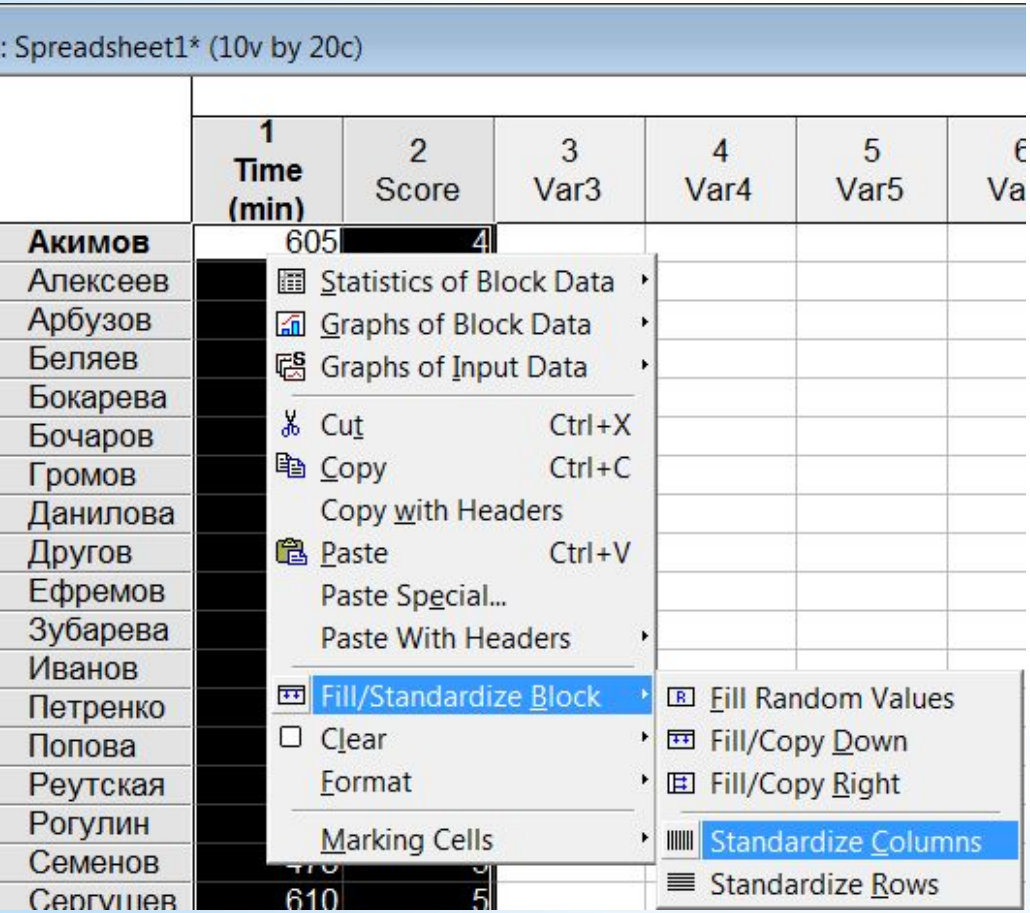
2. **Преобразуем каждое значение наблюдения по формуле:**

$$x_{st} = \frac{x_0 - \bar{x}}{\sqrt{\sigma^2}}$$

ИТОГ: мы получим значения переменных, колеблющиеся около нуля.

Добьёмся этого средствами STATISTICA 8.0 ->

Стандартизация переменных из контекстного меню. Получение новых значений

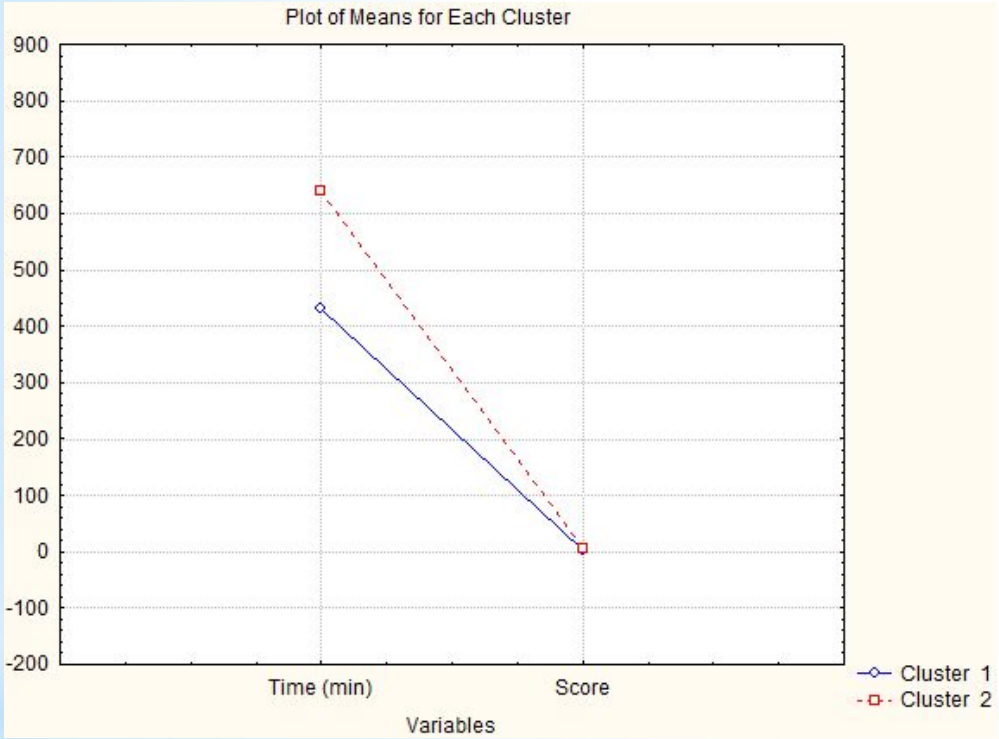


	1 Time (min)	2 Score
Акимов	0,514569	-0,17141
Алексеев	0,127675	-1,31414
Арбузов	-0,99002	-1,31414
Беляев	-1,60905	-1,31414
Бокарева	1,503298	0,971324
Бочаров	0,901463	-0,17141
Громов	0,832682	0,971324
Данилова	0,935853	-0,17141
Другов	-0,49136	-0,17141
Ефремов	-1,50588	-2,45688
Зубарева	0,299628	0,971324
Иванов	0,772498	0,971324
Петренко	-0,76648	-0,17141
Попова	1,374333	0,971324
Реутская	1,073416	0,971324
Рогулин	-0,73209	-0,17141
Семенов	-0,57733	0,971324
Сергушев	0,557557	0,971324
Смирнов	-0,97282	-0,17141
Степанов	-1,24795	-0,17141

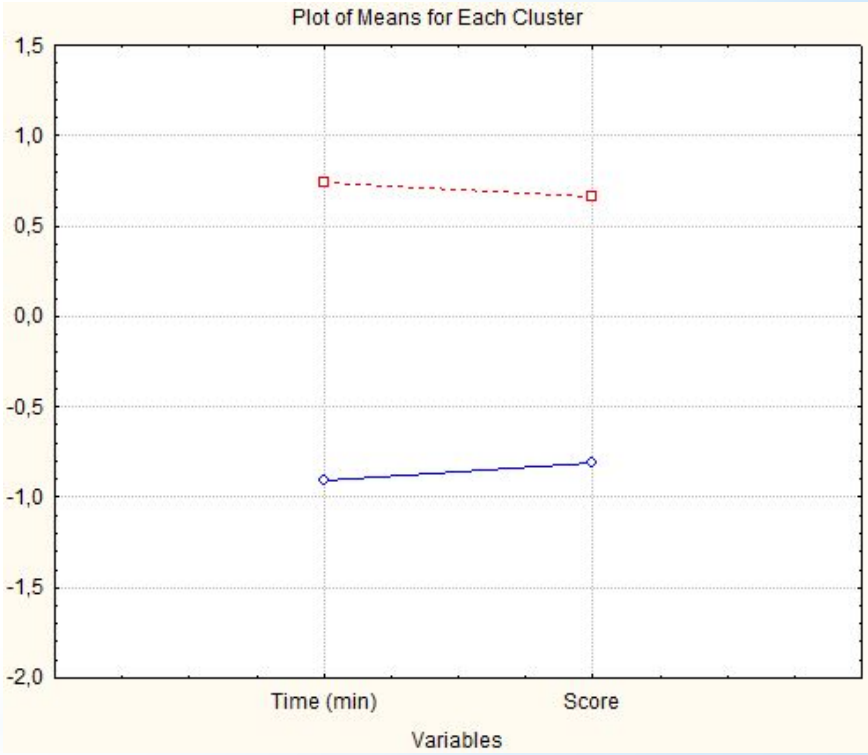
А теперь повторим процедуру кластерного анализа с «новыми» переменными…

Другое дело...

До стандартизации



После



Графики информируют нас о том, что студентов можно разбить на две группы, при этом первая группа характеризуется низкой посещаемостью класса (переменная «Time» равна -0,9097, т.е. время значительно ниже среднего) и низкими результатами на экзамене (переменная «Score» также существенно ниже средней и равна -0,8062).

Описательные статистики по кластеру

По Кластеру1

Descriptive Statistics for Cluster 1 (Spreadsheet1)

Variable	Descriptive Statistics for Cluster 1 (Spreadsheet1) Cluster contains 9 cases		
	Mean	Standard Deviation	Variance
Time (min)	-0,909774	0,532984	0,284072
Score	-0,806263	0,830177	0,689194

Евклидово расстояние между кластерами

Euclidean Distances between Clusters (Spreadsheet1)

Cluster Number	Euclidean Distances between Clusters (Spreadsheet1) Distances below diagonal Squared distances above diagonal	
	No. 1	No. 2
No. 1	0,000000	2,442558
No. 2	1,562869	0,000000

Цифры на картинке справа обозначают расстояния каждого объекта (в рассматриваемом примере — студента) до центра кластера. Поскольку центр кластера характеризует кластер, то чем меньше расстояния до центра, тем типичнее объект для данного кластера.

Поэлементный состав Кластера1

Members of Cluster Number 1 (Spreadsheet1)

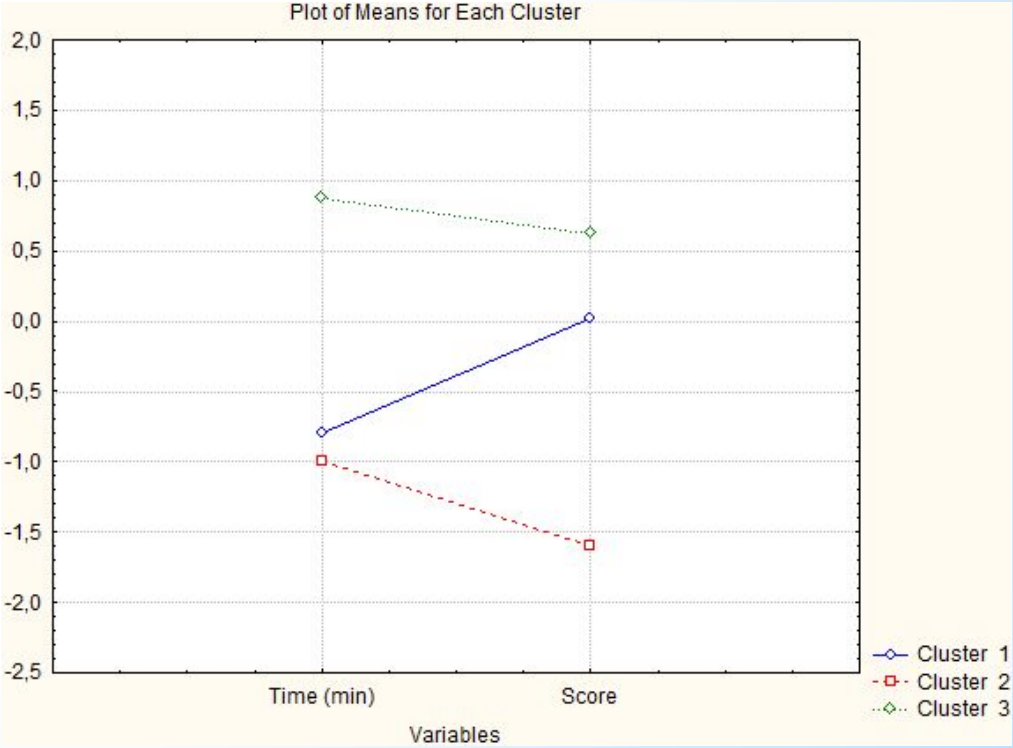
	Members of Cluster Number 1 (Spreadsheet1) and Distances from Respective Cluster Center Cluster contains 9 cases
	Distance
Алексеев	0,816775
Арбузов	0,363582
Беляев	0,611118
Другов	0,537639
Ефремов	1,240942
Петренко	0,460201
Рогулин	0,466160
Смирнов	0,451117
Степанов	0,508625

Больше кластеров – интереснее результаты?

Интерпретация

Выделяя три кластера, мы видим, что два из них весьма похожи на те кластеры, которых было только два. Смысл ***третьего кластера*** любопытен: фактически имеется **группа студентов, которые довольно вяло посещали дополнительные самостоятельные занятия, но получили средние**, а вовсе не плохие **оценки**.

Разбиение, число кластеров=3



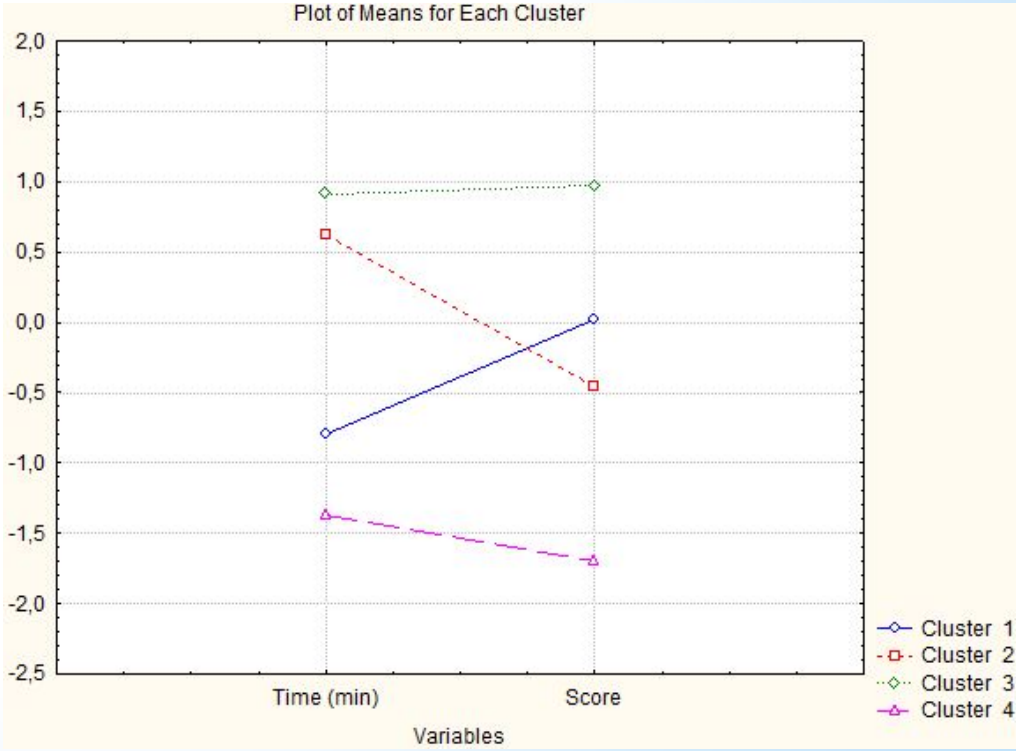
Вывод напрашивается сам собой: либо эти студенты вообще «продвинуты» в компьютерных технологиях и им на освоение нового программного продукта требуется гораздо меньше времени, либо они имеют изучаемые программы дома и работают с ними довольно много.

Больше кластеров – интереснее результаты?

Интерпретация

При разбивке на четыре кластера новый кластер обнаруживает группу студентов (в количестве 4 человек), которые, хотя и **усердно посещали компьютерный класс**, на экзамене **показали посредственные результаты**.

Разбиение, число кластеров=4



Descriptive Statistics for Cluster 2
Cluster contains 4 cases

Variable	Mean	Standard Deviation	Variance
Time (min)	0,619890	0,379686	0,144161
Score	-0,457094	0,571367	0,326460

Либо это просто слабые студенты, **либо** то, чем они занимались в компьютерном классе, имеет весьма отдаленное отношение к изучаемому предмету.

Особое значение проведенному анализу придает то, что мы можем выделить пофамильно студентов каждого кластера.

1. Суть кластерного анализа
2. История возникновения метода
3. Рассмотрение типичной задачи

   (с использованием STATISTICA 8.0)
4. **Методы кластерного анализа и его специфика**
5. Меры расстояния
6. Алгоритмы объединения в кластеры
7. Рассмотрение примера из сферы бизнеса

Алгоритм проведения кластерного анализа

Методы кластерного анализа относятся к так называемым **многомерным методам**. Перед исследователем находится поле из множества объектов, каждый из которых описывается **множеством переменных**.

Методы кластерного анализа позволяют разбить изучаемую совокупность объектов на группы объектов.

Кластерный анализ делится на несколько этапов.

1. **Спецификация проблемы**, т. е. выбор переменных, на основе которых будет производиться кластеризация.
2. **Выбор меры расстояния между объектами.**
3. **Преобразование переменных.**
4. **Выбор метода кластеризации.**
5. **Задание количества кластеров.**
6. **Интерпретация полученных результатов.**
7. **Оценка эффективности кластерного анализа.**

Методы кластерного анализа

1. АГГЛОМЕРАТИВНЫЕ

Исследователь начинает с создания элементарных кластеров, каждый из которых состоит только из одного исходного наблюдения (одной точки), а на каждом последующем шаге происходит объединение двух наиболее близких кластеров в один.

Графически процесс может быть представлен в виде дендрограммы, что позволяет видеть величину расстояния, на котором соответствующие элементы связываются в новый кластер.

2. ДИВИЗИВНЫЕ

Разбивка кластеров происходит непосредственно при заданном **заранее** числе кластеров. Метод К-средних строит ровно **К различных кластеров**, расположенных на возможно больших расстояниях друг от друга

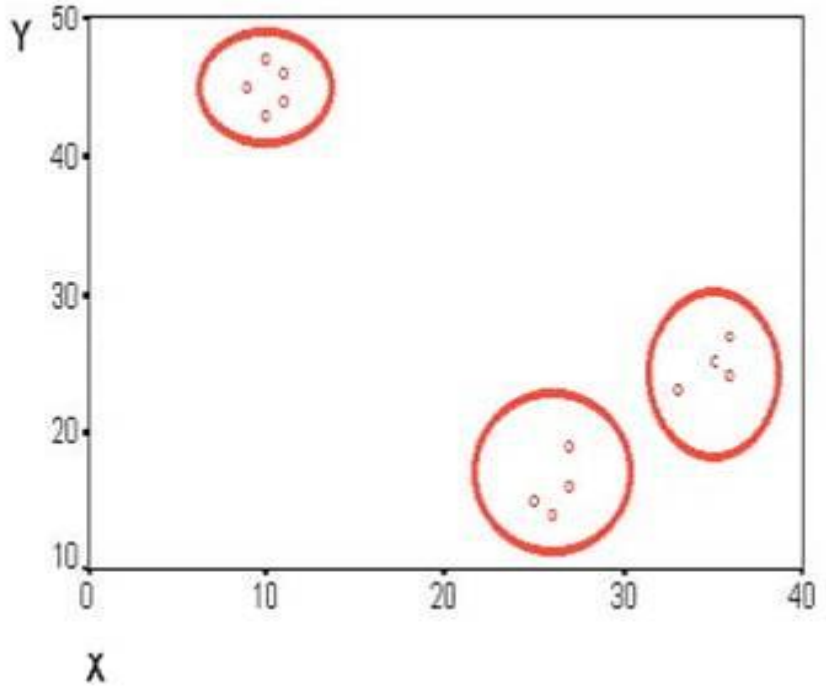
1. Суть кластерного анализа
2. История возникновения метода
3. Рассмотрение типичной задачи

   (с использованием STATISTICA 8.0)
4. Методы кластерного анализа и его специфика
5. **Меры расстояния**
6. Алгоритмы объединения в кластеры
7. Рассмотрение примера из сферы бизнеса

Меры расстояния

Для того чтобы определить близость, или схожесть, различных объектов, необходимо ввести некоторую количественную величину, характеризующую эту близость (схожесть). Естественным представляется ввести некоторую меру расстояния между объектами, аналогичную обычному физическому пространству.

Каждый объект будет представляться точкой в **многомерном пространстве признаков**. В таком случае кластеры будут выглядеть как скопления этих точек — своего рода «галактики» в «космическом пространстве».



Меры расстояния (1/2)

В кластерном анализе используют следующие меры для измерения расстояний.

1. Евклидово расстояние *(Euclidean distances)*. Наиболее общий тип расстояния. Хорошо известное из школьного курса как геометрическое расстояние. Вычисляется по формуле (по исходным, а не по стандартизованным данным):

$$\text{расстояние}(x,y) = [\Sigma_i (x_i - y_i)^2]^{1/2}$$

2. Квадрат евклидова расстояния *(Squared Euclidean distances)*. Применяется, чтобы придать большие веса более отдаленным друг от друга объектам:

$$\text{расстояние}(x,y) = \Sigma_i (x_i - y_i)^2$$

3. Расстояние городских кварталов *(City-block (Manhattan) distances)*. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

$$\text{расстояние}(x,y) = \Sigma_i |x_i - y_i|$$

Меры расстояния (2/2)

4. Расстояние Чебышева *(Chebychev ditances metric)*. Это расстояние может оказаться полезным, когда желают определить два объекта как "различные", если они различаются по какой-либо одной координате (каким-либо одним измерением).

$$\text{расстояние}(x,y) = \text{Максимум}|x_i - y_i|$$

5. Степенное расстояние. Иногда желают прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Это может быть достигнуто с использованием степенного расстояния:

$$\text{расстояние}(x,y) = (\Sigma_i|x_i - y_i|^p)^{1/r}$$

где *r* и *p* - параметры, определяемые пользователем. Если оба они равны 2, то это расстояние совпадает с расстоянием Евклида.

6. Процент несогласия *(Percent disagreement)*. Эта мера используется в тех случаях, когда данные являются категориальными.

$$\text{расстояние}(x,y) = (\text{Количество } x_i \neq y_i)/ i$$

1. Суть кластерного анализа
2. История возникновения метода
3. Рассмотрение типичной задачи

   (с использованием STATISTICA 8.0)
4. Методы кластерного анализа и его специфика
5. Меры расстояния
6. **Алгоритмы объединения в кластеры**
7. Рассмотрение примера из сферы бизнеса

Алгоритмы объединения в кластеры

На первом шаге мы измерили расстояния между нашими объектами, которые и рассматриваем в качестве первичных кластеров. Далее встаёт вопрос:

По какому правилу следует производить дальнейшее объединение?

Для этого также используется ряд методов.

1. Метод ближайшего соседа *(одиночная связь, Single linkage)*. Расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами («ближайшими соседями») в различных кластерах. Это правило похоже на «нанизывание» объектов для формирования кластеров, и результирующие кластеры имеют тенденцию быть представлены длинными «цепочками».

Алгоритмы объединения в кластеры

2. **Метод наиболее удаленного соседа** *(полная связь, Complete linkage)*. Расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах.

3. **Невзвешенное попарное среднее** *(Unweighted pair-group average)*. Расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них.

5. **Взвешенное попарное среднее** *(Weighted pair-group average).* Метод идентичен предыдущему за исключением того, что при вычислениях размер соответствующих кластеров (т. е. число содержащихся в них объектов) используется в качестве весового коэффициента. Поэтому предпочтительней использовать данный метод, если есть предположение о неравных размерах кластеров.

Алгоритмы объединения в кластеры

5. **Невзвешенный центроидный метод** (*Unweighted pair-group centroid)*. В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

6. **Взвешенный центроидный метод** *(медиана)*. Этот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них).

7. **Метод Варда** (*Ward's method).* Этот метод отличается от всех других методов, поскольку для оценки расстояний между кластерами он использует методы дисперсионного анализа. Метод **минимизирует сумму квадратов** для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

1. Суть кластерного анализа
2. История возникновения метода
3. Рассмотрение типичной задачи

   (с использованием STATISTICA 8.0)
4. Методы кластерного анализа и его специфика
5. Меры расстояния
6. Алгоритмы объединения в кластеры
7. **Рассмотрение примера из сферы бизнеса**

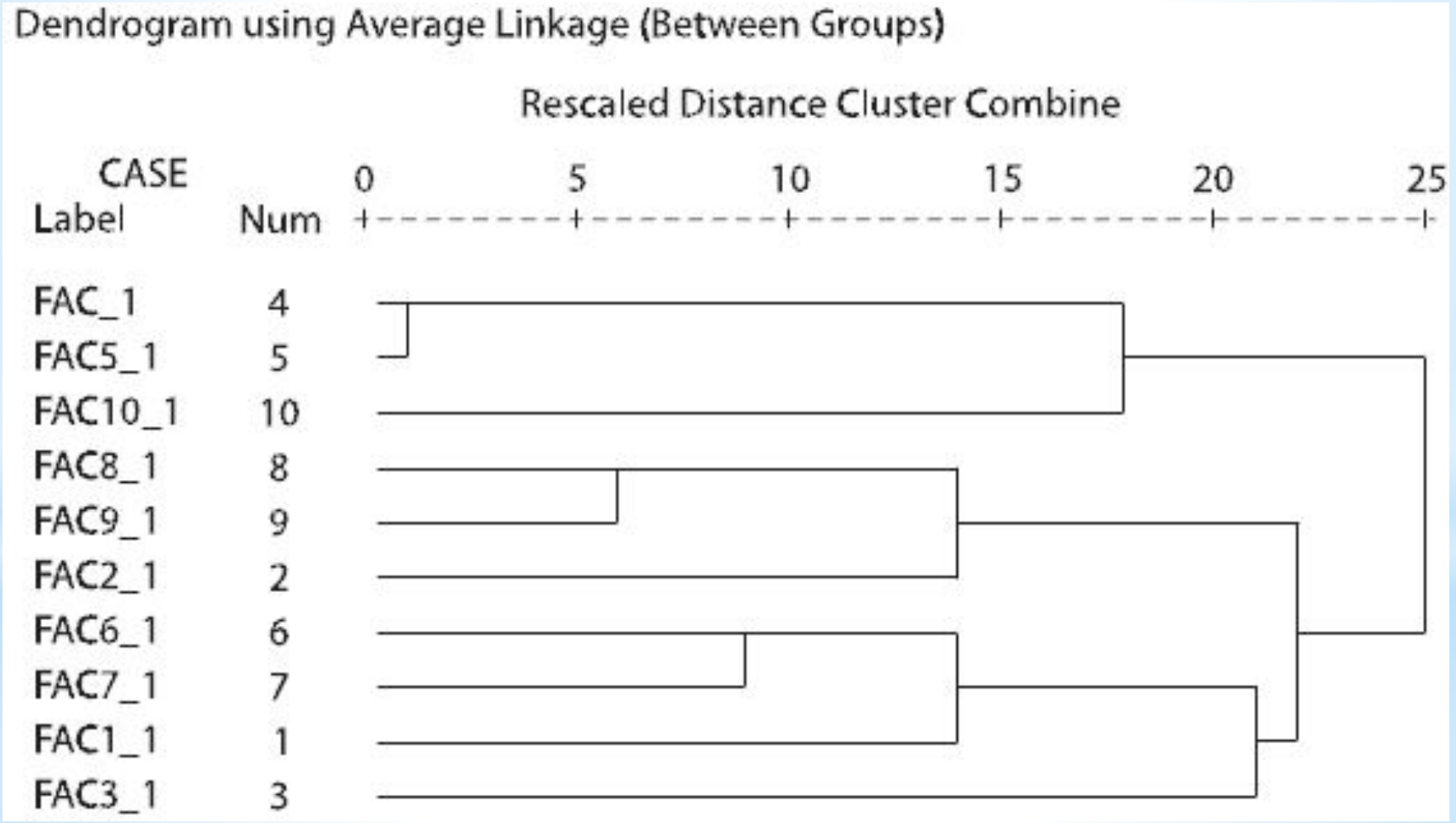
«КРИТИЧЕСКИЕ ФАКТОРЫ УСПЕХА ПРОЕКТА: НЕКОТОРЫЕ АСПЕКТЫ УПРАВЛЕНИЯ IT-ПРОЕКТАМИ В КИТАЕ»

Логика проведения анализа данных:

1. Анализ надёжности и достоверности
2. Факторный анализ
3. Кластерный анализ

Реальное исследование

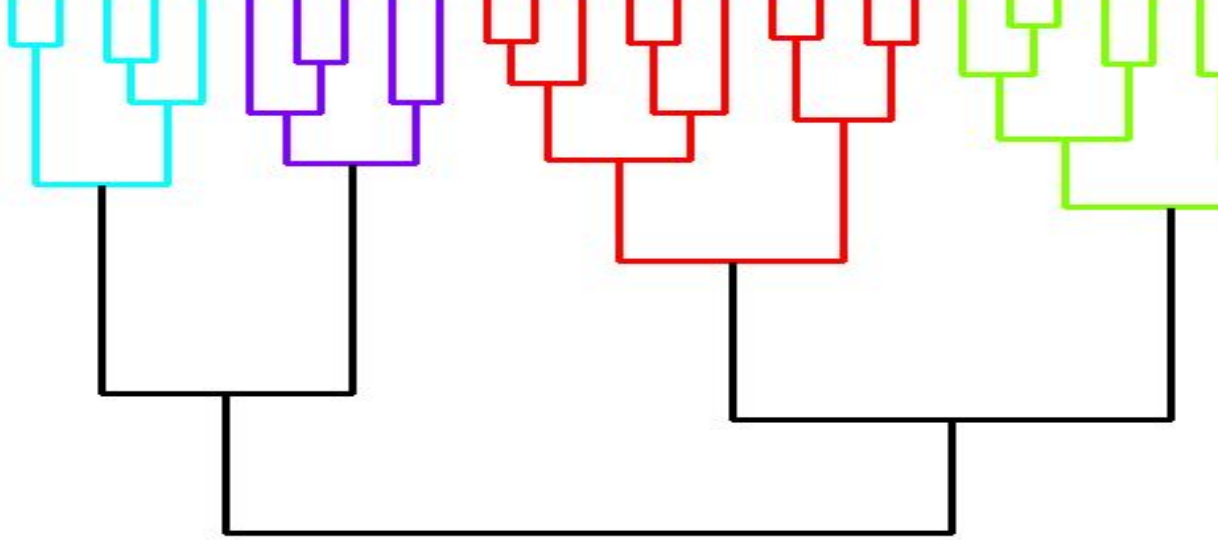
«КРИТИЧЕСКИЕ ФАКТОРЫ УСПЕХА ПРОЕКТА: НЕКОТОРЫЕ АСПЕКТЫ УПРАВЛЕНИЯ IT-ПРОЕКТАМИ В КИТАЕ»



Реальное исследование

«КРИТИЧЕСКИЕ ФАКТОРЫ УСПЕХА ПРОЕКТА: НЕКОТОРЫЕ АСПЕКТЫ УПРАВЛЕНИЯ IT-ПРОЕКТАМИ В КИТАЕ»

№ фактора	Название фактора	№ кластера	Название кластера
CSF3	Управление рисками	Кластер 3	Определение рисков
CSF1 CSF7 CSF6	Анализ требований Коммуникация и передача информации Постановка целей	Кластер 1	Анализ требований
CSF2 CSF8 CSF9	Механизм ограничений и мотивации Управление конфликтами Вклад участников	Кластер 4	Построение отношений
CSF4 CSF5 CSF10	Организация и координирование проекта Контроль процесса Окружение управления проектом	Кластер 2	Распределение ролей



Спасибо за внимание !