ЛАБОРАТОРНАЯ РАБОТА № 10
КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ предназначен для разбиения множества объектов на заданное или неизвестное число классов на основании некоторого критерия качества классификации (cluster — гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством). Критерий качества кластеризации отражает следующие неформальные требования:

1) внутри групп объекты должны быть тесно связаны между собой;
2) объекты разных групп должны быть далеки друг от друга;
3) при прочих равных условиях распределение объектов по группам должно быть равномерным.

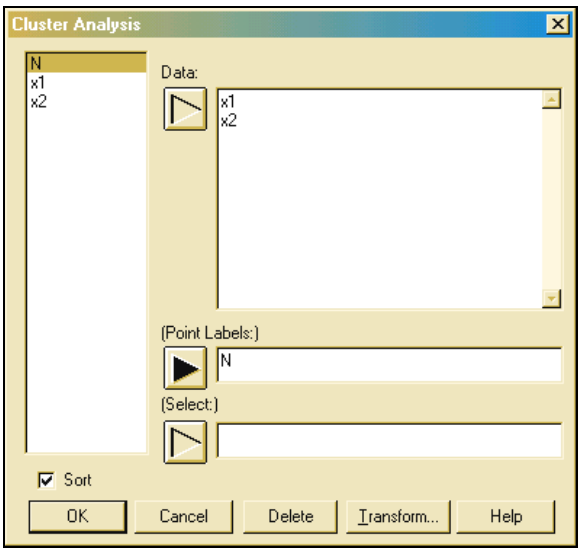
Алгоритмы кластерного анализа отличаются большим разнообразием. Широкое распространение получили алгоритмы иерархического группирования объектов и признаков, которые, в частности, достаточно полно представлены в *Statgraphics*. Эти алгоритмы предназначены для получения наглядного представления о стратификационной структуре всей исследуемой совокупности объектов. Они основаны на последовательном объединении кластеров (агломеративные процедуры) или на последовательном разбиении (дивизимные процедуры). Наибольшую популярность имеют агломеративные процедуры.

В Statgraphics реализовано 7 видов иерархических агломеративных процедур и одна неиерархическая процедура кластерного анализа.

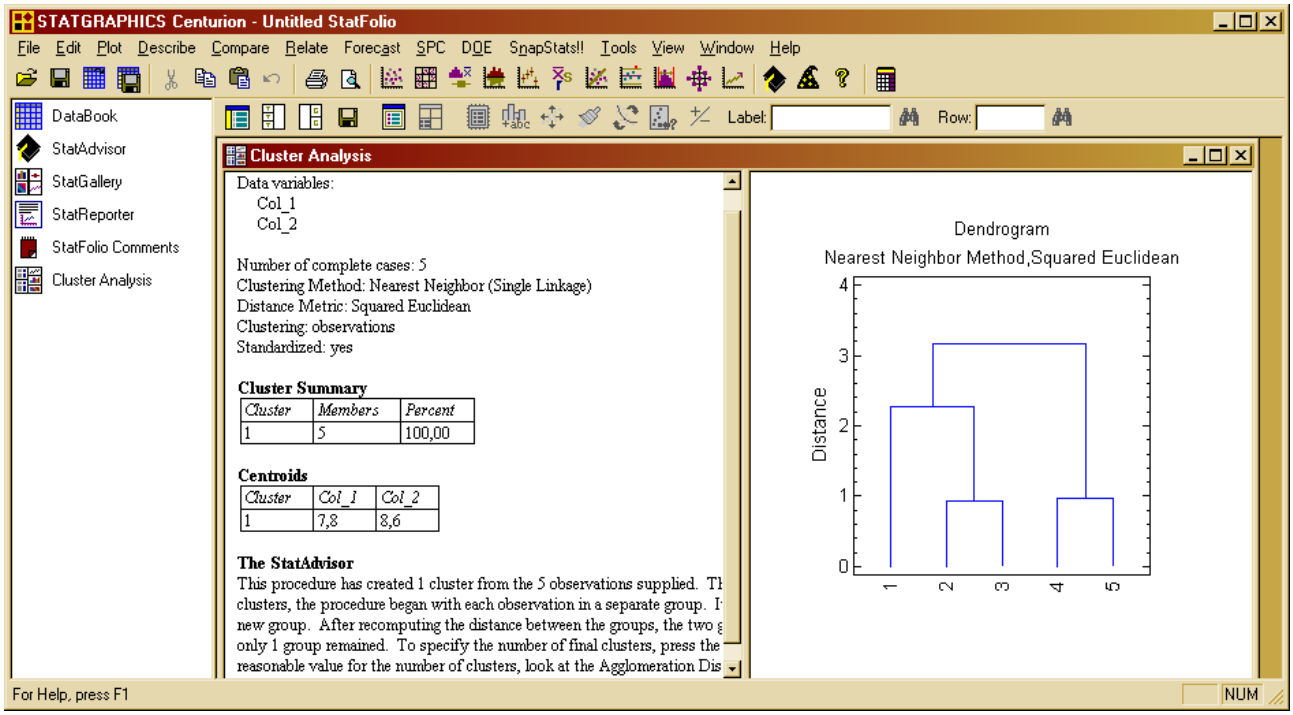
В качестве примера рассмотрим задачу: по данным, представленным в таблице, провести классификацию пяти семей по двум показателям: уровень расходов (млн руб.) за летние месяцы на культурные нужды, спорт и отдых (x_1) и питание (x_2).

№ семьи (i)	1	2	3	4	5
$x_i^{(1)}$	2	4	8	12	13
$x_i^{(2)}$	10	7	6	11	9

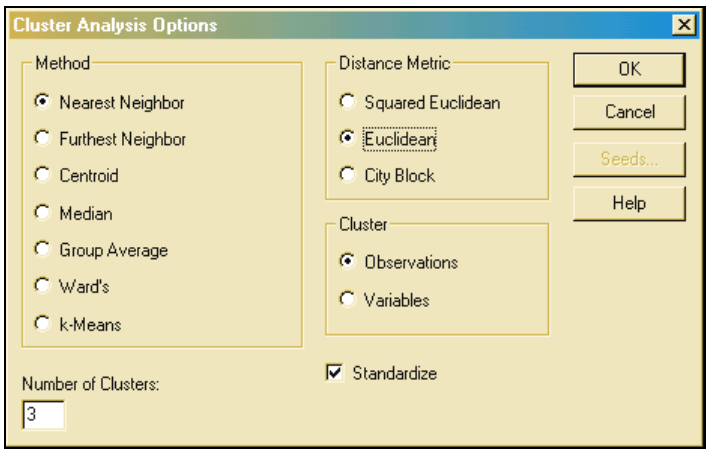
В меню выберите ***Describe***, в раскрывшемся меню выберите ***Multivariate Methosds***, затем ***Cluster Analysis***. Раскроется окно ***Cluster Analysis.***



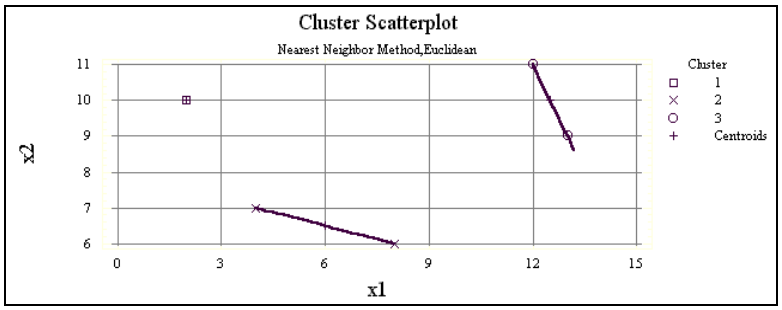
В поле ***Data*** надо поместить переменные *x*1 и *x*2 для задействования их в анализе. В поле ***Point Label*** поместите переменную *N*, нажмите кнопку ОК. Перед вами раскроется окно с первичной сводкой кластерного анализа.



По дендрограмме видно, что данные можно разбить на три кластера. Щелкните правой кнопкой мыши, в контекстном меню выберите ***Analysis Options***, раскроется окно ***Cluster Analysis Options:***



Выберите в поле ***Number of Clusters*** **3**, в ***Distance Metric*** выберите ***Euclidean***. В ***Method*** оставьте ***Nearest Neighbor*** (расстояние «ближний сосед»). Нажмите кнопку ОК, дендрограмма изменится. Построим теперь график. Нажмите кнопку ***Graphs***, выберите **2*D Scatterplot***, появится график. Для того чтобы лучше видеть кластеры, щелкните правой кнопкой, выберите ***Pane Options***, поставьте флажок против ***Circle Clusters***.



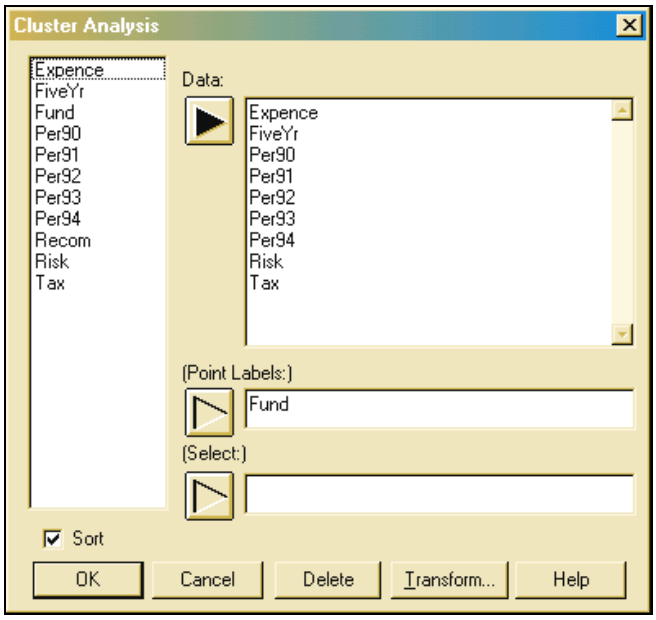
В данном случае вместо кругов прямые, так как в кластерах всего по две точки.

Попробуем теперь другой метод: объединим кластеры по принципу «дальнего соседа». Щелкните правой кнопкой мыши, в контекстном меню выберите ***Analysis Options***, раскроется окно ***Cluster Analysis Options, Method*** выберите ***Furthest Neighbor***.

Измените теперь самостоятельно количество кластеров, пусть их будет два. Самостоятельно попробуйте менять разные методы. Данные у нас измеряются в одинаковых единицах, поэтому необходимо отказаться от стандартизации данных. Щелкните правой кнопкой мыши, в контекстном меню выберите ***Analysis Options***, раскроется окно ***Cluster Analysis Options,*** снимите флажок в поле ***Standardize***.

Рассмотрим задачу о рынке ценных бумаг (проблему оценки различных фондов, оперирующих этими бумагами). Будем исследовать 16 известных инвестиционных фондов для оценки их состояния. В качестве переменных используются следующие характеристики: доходность за пятилетний период — *FiveYr*, *Risk* — риск, ежегодный процент дохода (для каждого года) — *Perf*90, *Perf*91, *Perf*92, *Perf*93, *Perf*94, расходная часть — переменная *Expence* и налоговые рейтинги — переменная *Tax*.

Данные для этого примера находятся в файле *Growth.sf*, откройте его. В меню выберите ***Describe***, в раскрывшемся меню выберите ***Multivariate Methods***, затем ***Cluster Analysis***. Раскроется окно ***Cluster Analysis.***

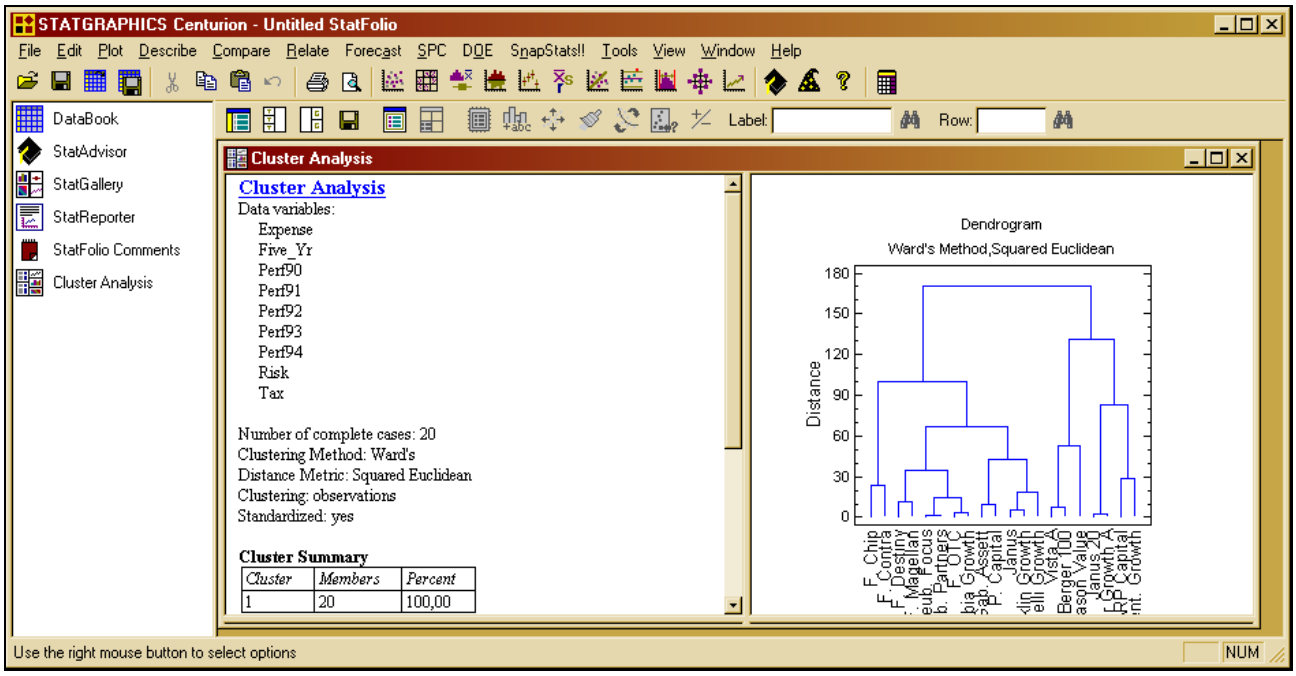


В поле ***Data*** надо поместить переменные ***Expence, FiveYr, Perf*90, *Perf*91, *Perf92, Perf93, Perf94, Risk*** и ***Tax*** для использования их в анализе. Для этого достаточно выделить нужные файлы и щелкнуть по стрелке в поле ***Data.*** В поле ***Point Label*** поместите переменную ***Fund***, нажмите кнопку OK. Перед вами раскроется окно с первичной сводкой кластерного анализа.

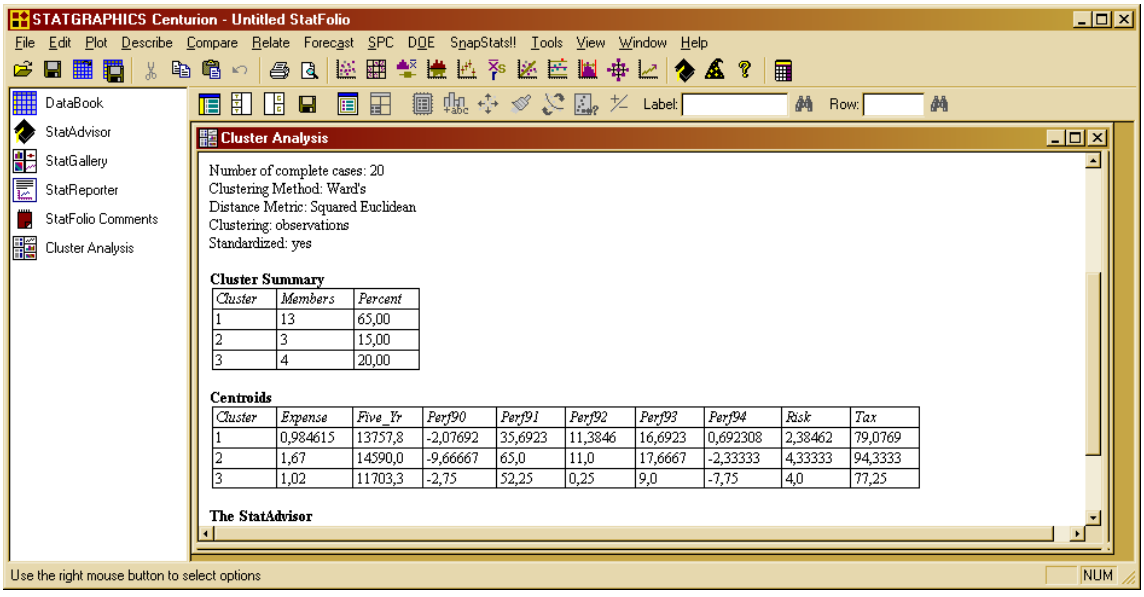
В нашем случае желательно, чтобы кластерный алгоритм хорошо работал с небольшим количеством наблюдений (у нас их всего 16) и был нацелен на выделение кластеров с приблизительно равным числом членов, остановим свой выбор на методе Варда (*Wards method*). Щелкните правой кнопкой по окну анализа, в меню выберите ***Analysis Options***, раскроется окно ***Cluster Analysis Op-***

tions, выберите в нем ***Ward's***. На экране отобразится сводка анализа для выбранного метода.

Посмотрим на графическое отображение результатов. Нажмите на кнопку ***Graphs***, выберите в появившемся окне ***Dendrogram***, нажмите кнопку ОК.



Дендрограмма отображает иерархическую структуру группирования инвестиционных фондов. Для более подробного рассмотрения группировок зададим количество кластеров, равным 3. Вызовите снова окно ***Cluster Analysis Options***, в поле ***Number of Clusters*** поставьте цифру 3. Нажмите ОК. Рассмотрим теперь внимательнее сводку результатов кластерного анализа. Щелкните дважды по соответствующему окну.



В сводке кластерного анализа прежде всего указываются имена переменных, участвующих в анализе, количество полных образцов (в нашем случае — 20), используемый метод кластерного анализа (в нашем случае — метод Варда), принятая метрика (у нас — Евклидова). Затем в сводке описывается число кластеров, количество объектов (*Members*), соответствующий процент населенности (*Percent*).

Далее идет информация о центроидах. По координатам центроидов можно судить о том, какие переменные играют наиболее важную роль в каждом кластере. В первом кластере видно, что расходы были разумными: несмотря на малые доходы в 1990 году, заметно, что в других годах состояние фондов первого кластера постоянно улучшалось. Также в первом кластере самый низкий рейтинг риска среди всех кластеров (*Risk*), а налоговые сборы были тоже достаточно невелики (*Tax*).

Переменные, представляющие второй кластер, говорят о том, что здесь имелись наибольшие расходы, хотя за пятилетний период доходы оставались самыми высокими. Оценка риска и налоговые сборы являются максимальными среди всех кластеров.

О третьем кластере можно сказать, что он занимает второе место по расходам относительно к доходам за пятилетний период. Оценка риска была самая высокая, однако налоговые сборы существенно ниже, чем у первого кластера.

Нажмите на кнопку ***Tables***, в раскрывшемся окне выберите ***Membership Table***, раскроется таблица принадлежности наблюдений. В данной таблице описаны выбранные параметры кластерного анализа, а затем дается полный список всех наблюдений, их имена и номера кластеров, в которые входят указанные наблюдения.

Посмотрим опять на графическое представление результатов. Нажмите на кнопку ***Graphs***, в раскрывшемся окне выберите **2*D Scatterplot*** (двумерная диаграмма рассеивания). Раскроется окно с графиком.

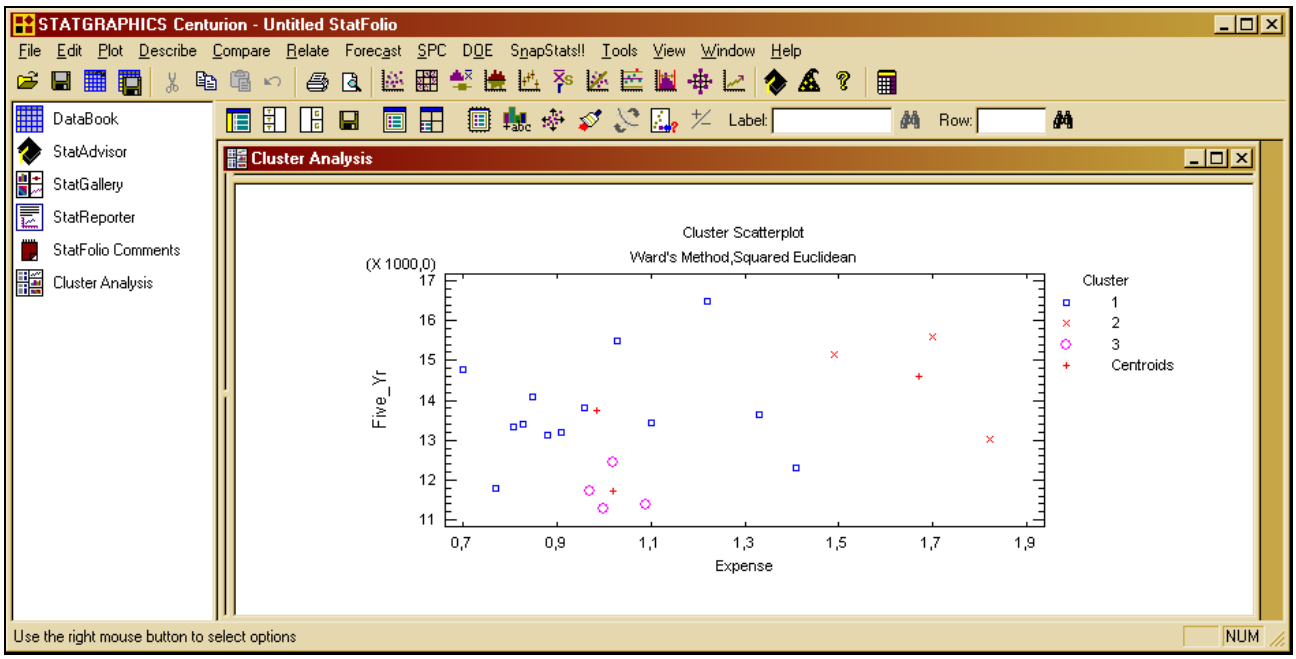


Диаграмма рассеивания показывает, как группируются исследуемые наблюдения на плоскости двух переменных ***Expence*** (расходная часть) и ***Five_Yr*** (доходность за пятилетний период). Каждый кластер представлен на диаграмме собственным символом и своим цветом (на экране). На легенде в правой части диаграммы представлены обозначения кластеров, кстати их можно менять, вспоминайте, как это сделать.

Из графика видно, что первый кластер имеет низкие относительные расходы. Во втором кластере наблюдаются самые высокие расходы, но и самые высокие максимальные доходы. В третьем кластере низкие расходы сопровождаются и невысокими пятилетними доходами.

Попробуйте изменять самостоятельно количество кластеров, проанализируйте полученные результаты. Покажите результаты преподавателю.

ЗАДАНИЯ

1. В таблице (а также в файле *Cluster.sf*) представлены значения следующих p=6 показателей, характеризующих условия жизни населения 20 стран в 1994 году:

$x^{(1)}$ — потребление мяса и мясопродуктов на душу населения (кг);

$x^{(2)}$ — смертность населения по причине болезни органов кровообращения на 100000 населения;

$x^{(3)}$ — оценка валового внутреннего продукта по паритету покупательной способности в 1994 году на душу населения (в % к США);

$x^{(4)}$ — расходы на здравоохранение (в % от ВВП);

$x^{(5)}$ — потребление фруктов и ягод на душу населения (кг);

$x^{(6)}$ — потребление хлебопродуктов на душу населения (кг).

№ п/п	Страны	Показатели					
		$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$
1	Россия	55	84,98	20,4	3,2	28	124
2	Австралия	100	30,58	71,4	8,5	121	87
3	Австрия	93	38,42	78,7	9,2	146	74
4	Азербайджан	20	60,34	12,1	3,3	52	141
5	Армения	20	60,22	10,9	3,2	72	134
6	Белоруссия	72	60,79	20,4	5,4	38	120
7	Бельгия	85	29,82	79,7	8,3	83	72
8	Болгария	65	70,57	17,3	5,4	92	156
9	Великобритания	67	34,51	69,7	7,1	91	91
10	Венгрия	73	64,73	24,5	6,0	73	106
11	Германия	88	36,63	76,2	8,6	138	73
12	Греция	83	32,84	44,4	5,7	99	108
13	Грузия	21	62,64	11,3	3,5	55	140
14	Дания	98	34,07	79,2	6,7	87	102
15	Ирландия	99	39,27	57,0	6,7	87	102
16	Испания	89	28,46	54,8	7,3	103	72
17	Италия	84	30,27	72,1	8,5	169	118
18	Казахстан	61	69,04	13,4	3,3	10	191
19	Канада	98	25,42	79,9	10,2	123	77
20	Киргизия	46	53,13	11,2	3,4	20	134

Требуется провести классификацию стран по уровню жизни населения и дать содержательную интерпретацию полученным результатам.

2. В 1999 году для поступления на математико-механический факультет было проведено весеннее тестирование и олимпиада. Результаты тестирования и олимпиады у одних и тех же людей приведены в файле *Olimp.sf*. Проведите кластерный анализ результатов, разбейте все результаты на 4 кластера (отличники, «хорошисты», «троечники», «двоечники»). Попробуйте использовать различные методы и метрики.

Пусть по результатам этих испытаний вы можете зачислить на факультет 82 человека. Проведите «зачисление» по результатам вашего анализа.

3. В файле с данными о рынке жилья проведите кластерный анализ квартир, разбейте квартиры на группы по цене.

ВОПРОСЫ

1. В каких случаях в качестве меры близости между объектами используется обычное евклидово расстояние, а в каких — нормализованное евклидово?
2. Для каких признаков используется Хеммингово расстояние?
3. Что можно использовать в качестве расстояния между признаками (не объектами)?
4. Как записывается расстояние между двумя кластерами по принципу «ближнего соседа»?
5. Как записывается расстояние между двумя кластерами по принципу «дальнего соседа»?
6. Как записывается расстояние между двумя кластерами с использованием расстояния «по центрам тяжести»?