ТЕМА 9

КЛАСТЕРНЫЙ АНАЛИЗ

Лабораторные работы 17 и 18

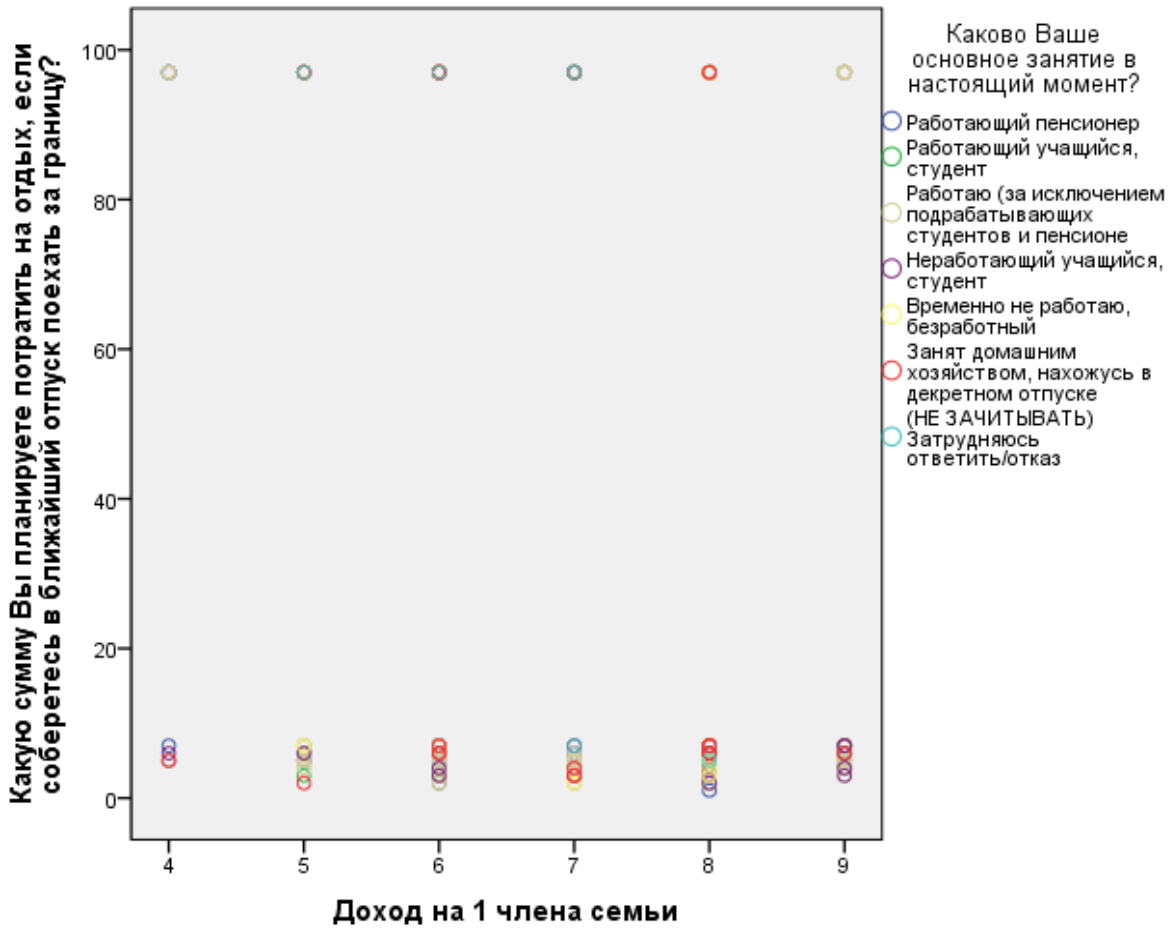
Лабораторная работа №17

Кластерный анализ

Лабораторная работа №17

1. Откройте массив данных leisure&tourism.sav, который содержит отношение респондента к отпуску и отдыху.
2. Возьмём переменные Q6 и INCOME и представим их при помощи простой **Диаграммы рассеяния**.
    - Выберите в меню «**Графики**» → «**Диаграмма рассеяния».**
    - Переменную INCOME поместите в поле оси x, а переменную Q6 в поле оси y, и для обозначения наблюдения используйте переменную PROF (основное занятие респондента).
    - Через кнопку «**Опции»** активируйте опцию «**Показывать график с метками наблюдений»**.

Лабораторная работа №17



- Следовательно, переменные INCOME и Q6 распадаются на различные кластеры по основному занятию респондента.

Лабораторная работа №17

- Самой распространенной мерой для определения расстояния между двумя точками на плоскости, образованной координатными осями x и y, является евклидова мера:

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

- В соответствии с этой формулой расстояние между основным занятием респондента составляет:

$$\sqrt{(90 - 8)^2 + (90 - 8)^2} = 164$$

$$\sqrt{7^2 + (90 - 7)^2} = 83{,}3$$

$$83{,}3 + 164 = 247{,}3$$

- К сожалению, интерпретировать картину отношений между переменными не так легко. Во-первых, структуры кластеров, если вообще таковые имеются, не так чётко разделены, особенно при наличии большого количества наблюдений. Скорее наоборот, кластеры размыты и даже проникают друг в друга. Во-вторых, как правило, кластерный анализ проводится не с двумя, а с намного большим количеством переменных.

Лабораторная работа №17

2. Команда **«Анализ»** → **«Классификация»** → **«Иерархический кластерный анализ»**. Без графиков, только «**Протокол объединения объектов в иерархическом кластерном анализе»**

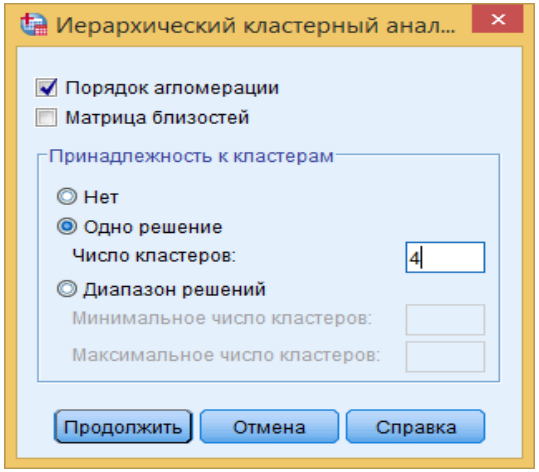
Порядок агломерации (кластеров)

Этап	Объединенный кластер		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	497	500	,000	0	0	4
2	478	499	,000	0	0	23
3	491	498	,000	0	0	10
4	22	497	,000	0	1	7
5	483	496	,000	0	0	18
6	494	495	,000	0	0	7
7	22	494	,000	4	6	16
8	460	493	,000	0	0	41
9	392	492	,000	0	0	108
10	13	491	,000	0	3	25
11	476	490	,000	0	0	25
12	261	489	,000	0	0	238
13	439	488	,000	0	0	62
14	432	487	,000	0	0	69
15	485	486	,000	0	0	16
16	22	485	,000	7	15	28

- Для определения числа кластеров - смотреть на значения коэффициентов (расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры (здесь: квадрат евклидового расстояния). На том этапе, где эта мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить.
- В приведенном примере — это скачок с 11,031 до 8447,426. То есть после не следует производить никаких последующих объединений.

Лабораторная работа №17

- После определения оптимального количества кластеров – можно установить принадлежность каждого объекта определенному кластеру.
- **«Иерархический кластерный анализ» - «Статистики» – «Принадлежность к кластерам» – «Одно решение»** - желаемое количество кластеров 4.
- Информацию о принадлежности каждого наблюдения к определённому кластеру можно сохранить в новой переменной: **«Сохранить» – «Одно решение» - «4».** Теперь для каждого наблюдения будет выводиться и информация о принадлежности к кластеру.
- **Дендрограмма** – графическое изображение объединения. Иногда она помогает визуально определить число кластеров.
- Как правило, в качестве кластерообразующих признаков не используют социально-демографические характеристики. Их используют для изучения и сравнения уже полученных кластеров.



Лабораторная работа №17

- В таблице «Принадлежность к кластерам» можно увидеть разбиение наблюдений на 4 кластера.
- Отметим, что иерархический кластерный анализ эффективен для малого количества объектов. Не годится для массивов большого объема из-за трудности агломеративного алгоритма и практической бессмысленности дендрограмм. В такой ситуации целесообразно использовать **алгоритм k-средних**.

Принадлежность к кластерам

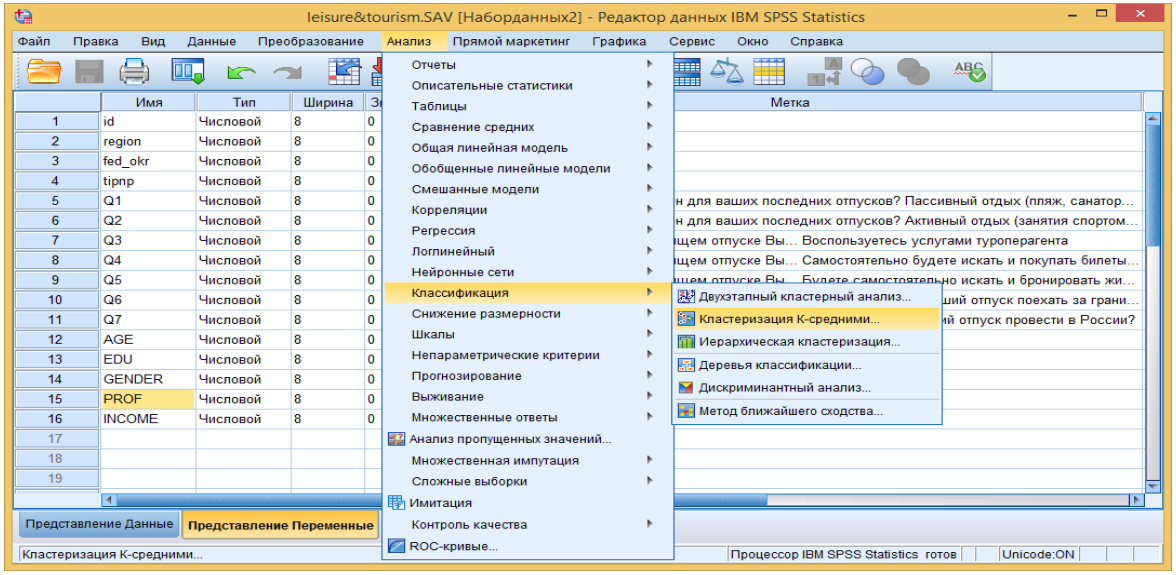
Наблюдение	Кластеры 4
1	1
2	2
3	1
4	1
5	2
6	1
7	2
8	2
9	2
10	2
11	2
12	2
13	3

Лабораторная работа №18

Кластеризация к-средними

Лабораторная работа №18

- Откройте массив данных leisure&tourism.sav
- Команда **«Анализ»** → **«Классификация»** → **«Кластеризация К-средними»**
- Остановимся также на 4 кластерах
- Лучше указывать больше итераций, например, 99
- Можно сохранить в качестве новых переменных номер кластера и расстояние каждого респондента от центра кластера. (команда «**Сохранить»**)
- Сравнение средних расстояний показывает степень однородности каждого кластера



Лабораторная работа №18

В результате мы получаем 4 кластера, с указанием количества наблюдений, отнесенных в каждый кластер.

Число наблюдений в каждом кластере

Кластеризовать	1	214,000
	2	74,000
	3	100,000
	4	112,000
Допустимо		500,000
Пропущенные		,000