

Практическое занятие № 1

Преобразование и фильтрация данных

С помощью инструментов предварительной обработки данных в Deductor Studio можно добиться решения ряда промежуточных аналитических задач по обогащению данных. Так, при проведении анализа или построении модели прогноза часто приходится разбивать исходные данные на группы, исходя из определенных критериев.

В первом случае такая необходимость возникает, если аналитик желает просмотреть, к примеру, информацию не по всей совокупности данных, а по определенным группам (например, какую сумму кредита берут заёмщики на те или иные цели, либо к какой возрастной категории они относятся).

Во втором случае (при прогнозировании) аналитику необходимо учитывать тот факт, что определенные группы (например, категории заемщиков) ведут себя по-разному, и что модель прогноза, построенная на всех данных не будет учитывать нюансов, возникающих в этих группах. Т.е. лучше построить несколько моделей прогноза, например, в зависимости от суммовой группы кредита, и строить прогноз на них, нежели построить одну модель прогноза. Исходя из этого и не только, в Deductor Studio предоставляется широкий набор инструментов, тем или иным способом позволяющих разбивать исходные данные на группы, группировать любым способом всевозможные показатели и т.п.

Решение задачи по преобразованию и фильтрации данных рассмотрим на примере данных по рискам кредитования физических лиц.

Исходные данные: текстовый файл **Credit.txt**, хранящийся на сетевом диске в папке **!Tasks**. Содержит таблицу с рядом полей, среди которых: «Сумма кредита», «Дата кредитования», «Цель кредитования» и «Возраст» кредитора.

Требуется: реализовать разбиение на группы и фильтрацию табличных данных и исследовать их взаимовлияние с помощью визуального представления «Куб».

1. Разбиение данных на группы и фильтрация.

1. Запустить пакет Deductor Studio Academic, версия 5.2 (или иная).
2. Инициировать Мастер импорта с целью импортирования в среду пакета текстового файла Credit.txt.
3. На шаге 6 Мастера импорта установить назначение полей текстового файла: *Сумма кредита* – факт, *Цель кредитования* и *Возраст* – измерение, остальные поля – параметры по умолчанию (рис. 1).
4. Продолжить операции импорта. На шаге 8 Мастера импорта установить флажки отображения результатов импорта в виде таблицы и куба (рис.2).
5. На шаге 9 Мастера импорта убедиться в правильности задания назначения полей *Сумма кредита* – *факт*, *Цель кредитования* и *Возраст* – *измерение*. Для остальных полей таблицы установить назначение – *неиспользуемое* (рис. 3).

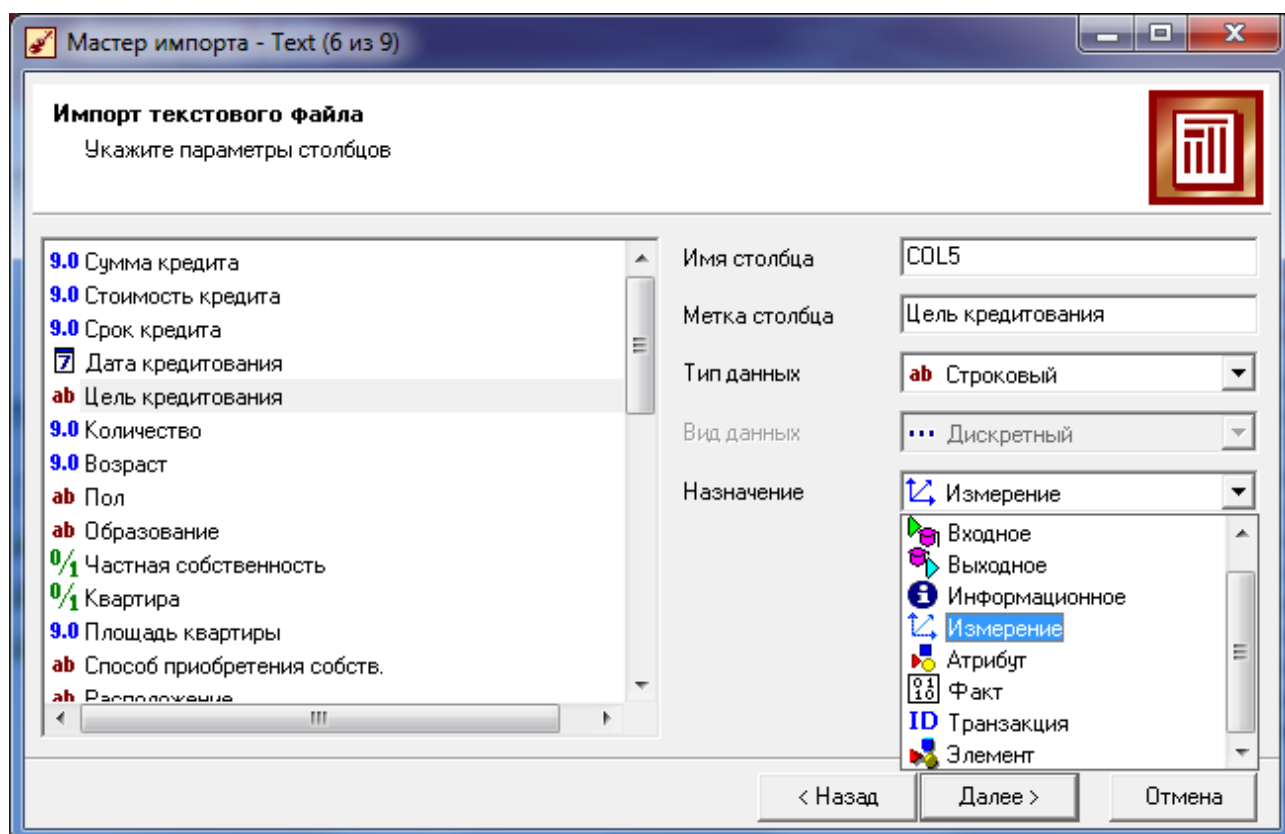


Рис. 1. Окно задания параметров импорта с раскрытым списком назначения для столбца Цель кредитования

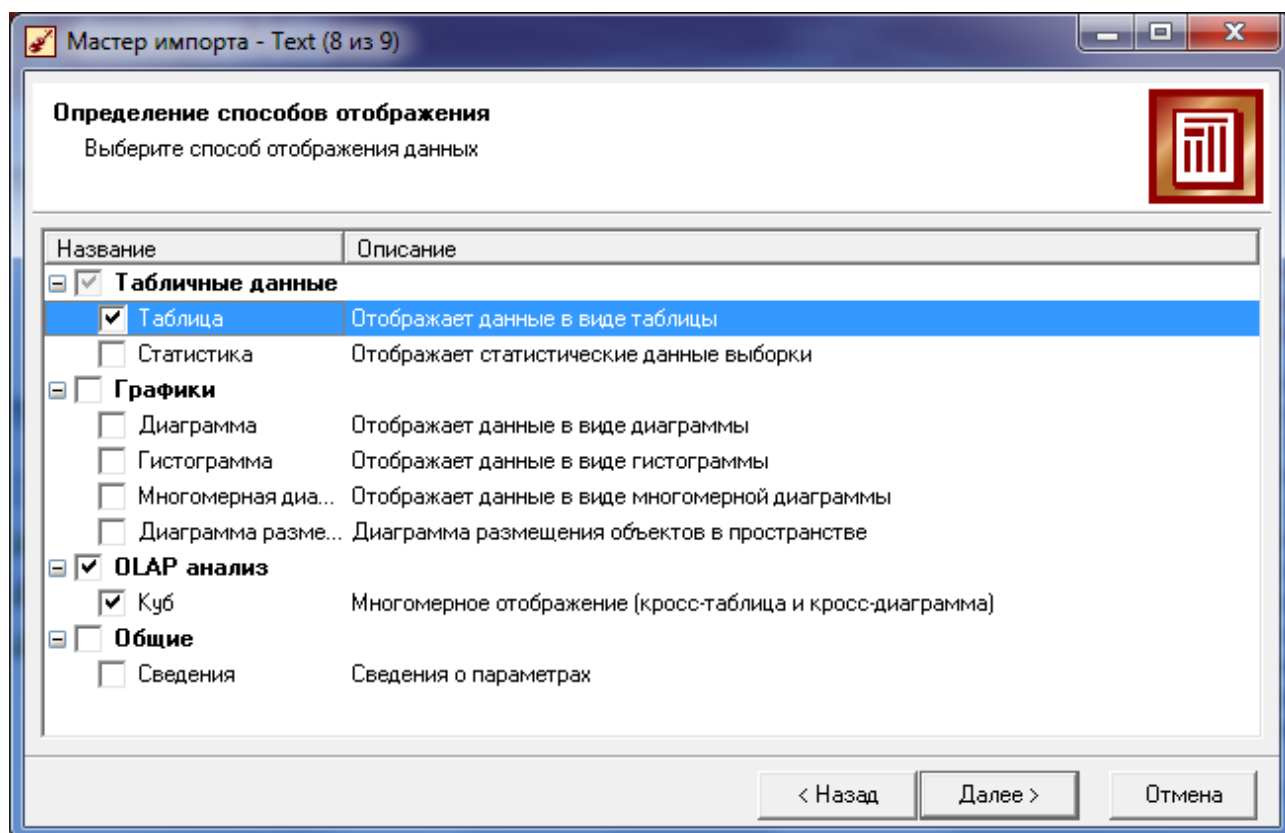


Рис. 2. Задание способов отображения результатов импорта

6. Продолжить выполнение импорта данных. На шаге 10 Мастера импорта настроить размещение измерений – Цель кредитования – *строки*, Возраст – *столбцы*,

перетаскивая их с помощью мыши в соответствующие окна из области доступных измерений (рис. 4).

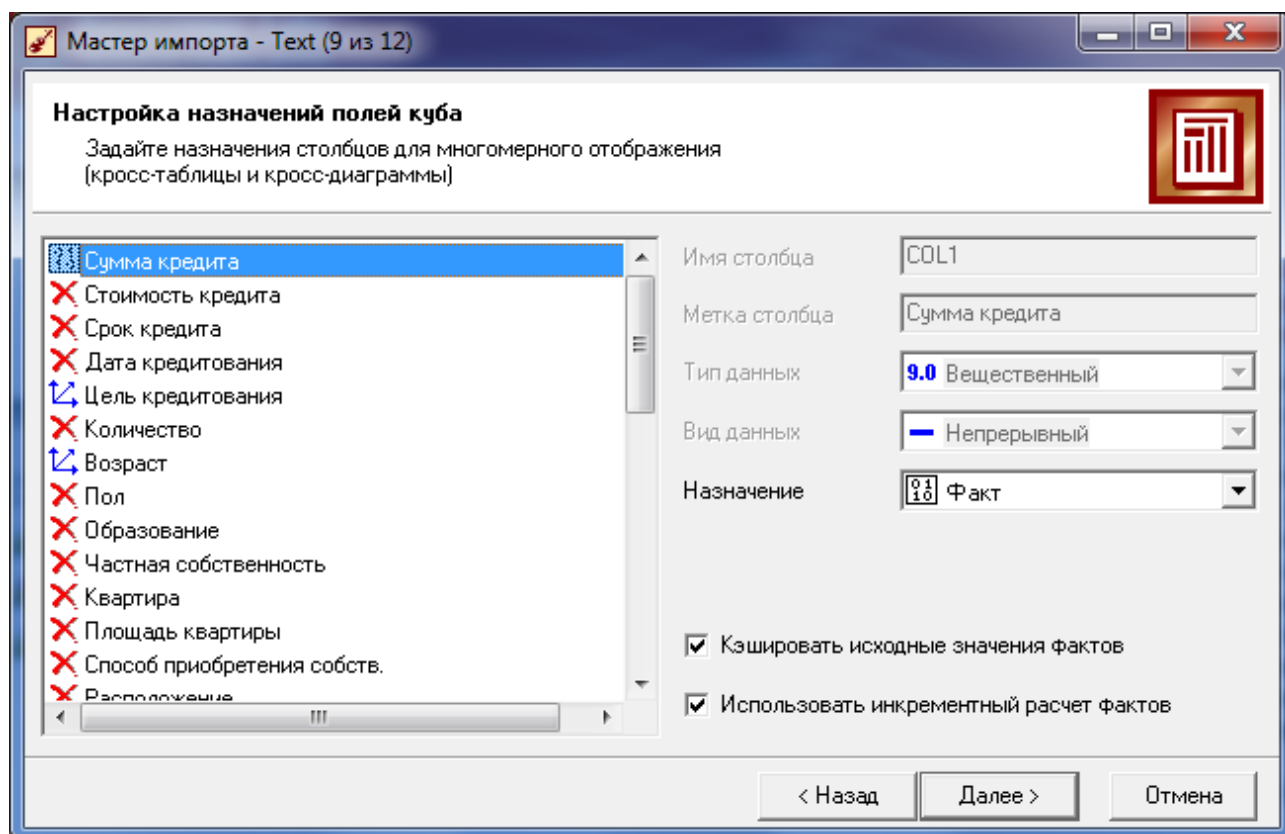


Рис. 3. Задание назначений полей куба

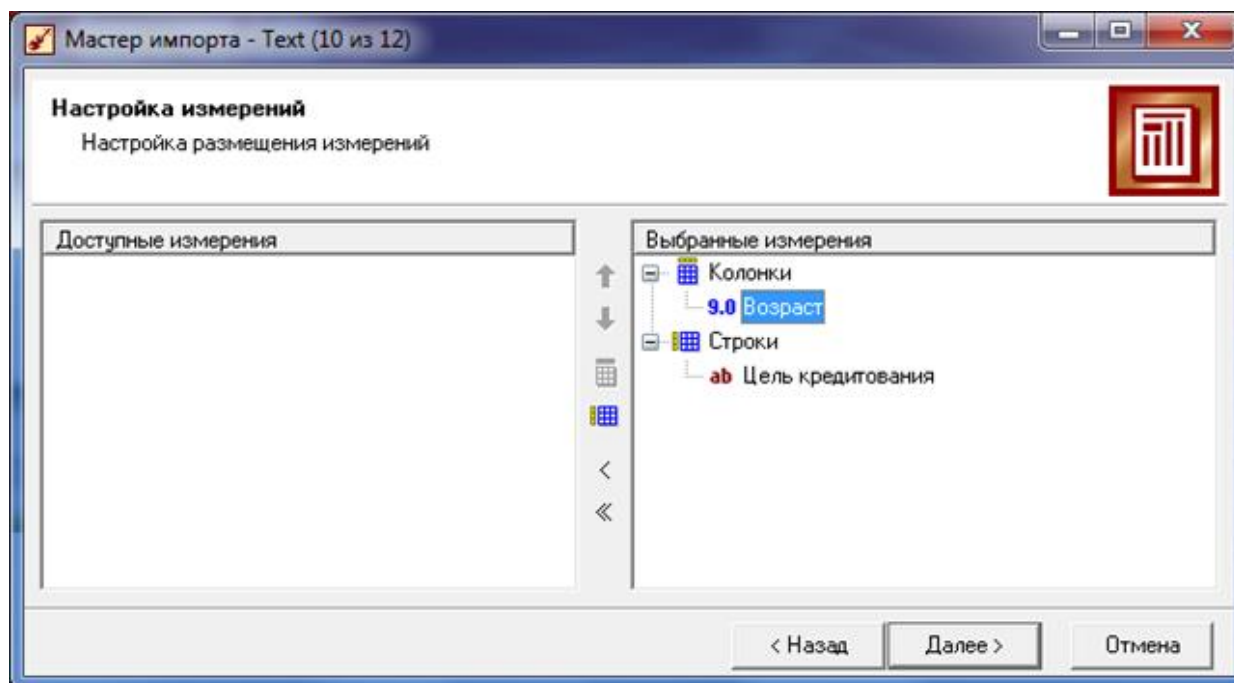


Рис. 4. Настройка измерений куба

7. На шаге 11 Мастера импорта установить флажок *факта* для столбца *Сумма кредита*.
8. Перейти к завершению импортирования.

9. Выполнить анализ результатов импортирования – в виде таблицы (рис. 5) и в виде куба (рис. 6).

Таблица

1 / 149

Сумма кредита	Стоимость кредита	Срок кредита	Дата кредитования	Цель кредитования
7000	1400	6	01.01.2003	Иное
7500	1500	6	01.01.2003	Иное
14500	2900	12	01.01.2003	Покупка товара
15000	3000	6	01.01.2003	Покупка товара
32000	6400	12	01.01.2003	Иное
11500	2300	6	01.01.2003	Турпоездки, развлечения и т.п.
5000	1000	6	01.01.2003	Покупка и ремонт недвижимости
61500	12300	30	01.01.2003	Покупка товара
13500	2700	12	01.01.2003	Оплата услуг (мед., юрид. и т.п.)
25000	5000	18	01.01.2003	Покупка товара
25500	5100	24	01.01.2003	Покупка товара
9500	1900	6	01.01.2003	Покупка товара
53000	10600	24	01.01.2003	Иное
27500	5500	18	02.01.2003	Покупка товара
4000	800	6	02.01.2003	Оплата услуг (мед., юрид. и т.п.)
40500	8100	24	02.01.2003	Покупка и ремонт недвижимости
51500	10300	36	02.01.2003	Покупка и ремонт недвижимости
7000	1400	6	02.01.2003	Оплата услуг (мед., юрид. и т.п.)

Рис. 5. Результат импортирования в виде таблицы

Куб

Возраст

Цель кредитования	19	20	21	22	23	24	25	26	27	28	29	30
Иное	50 000,00	17 000,00	8 500,00	23 500,00			87 000,00			45 500,00		59 000,00
Оплата за образование		17 500,00	29 500,00	31 500,00	49 000,00	23 500,00	18 500,00	34 500,00	73 500,00	19 500,00		
Оплата услуг (мед., юрид. и т.п.)					66 500,00	25 000,00		9 000,00		6 500,00		37 000,00
Покупка и ремонт недвижимости	78 000,00		13 000,00	46 500,00	27 500,00	9 500,00	64 500,00		38 000,00	14 500,00	13 500,00	23 500,00
Покупка товара	46 500,00	73 500,00	76 500,00	112 000,00	98 500,00	61 500,00	55 500,00			15 000,00	61 500,00	95 000,00
Турпоездки, развлечения и т.п.		30 500,00					11 500,00		8 500,00		15 500,00	
Итого:	174 500,00	138 500,00	127 500,00	213 500,00	241 500,00	119 500,00	237 000,00	43 500,00	120 000,00	101 000,00	90 500,00	214 500,00

Рис. 6. Результат импортирования в виде куба

10. Используя кнопку *Транспонирование* – смена местами строк и столбцов, изменить расположение измерений (рис. 7).
11. Обратите внимание, что для наименований измерений - *Возраст* и *Цель кредитования*, можно задать параметры *фильтрации*, раскрыв список их значений и установив соответствующие флажки. Для списка значений измерения *Цель кредитования*, используя кнопку *Выделение по маске*, задать текст *Содержит значение Оплата* (рис. 8).
12. Завершить задание фильтра кнопками ОК и ☒ (ОК). Проанализировать результат фильтрации.
13. С помощью кнопки *Отображать кросс-диаграмму*, разместить ее внизу таблицы. Задать для представления легенды диаграммы место внизу диаграммы (рис. 9).

Куб

Цель кредитования: Транспонирование - смена местами строк и столбцов (Ctrl+T)

Возраст	Иное	Оплата за с	Оплата усл	Покупка и р	Покупка то	Турпоездки	Итого:
19	50 000,00			78 000,00	46 500,00		174 500,00
20	17 000,00	17 500,00			73 500,00	30 500,00	138 500,00
21	8 500,00	29 500,00		13 000,00	76 500,00		127 500,00
22	23 500,00	31 500,00		46 500,00	112 000,00		213 500,00
23		49 000,00	66 500,00	27 500,00	98 500,00		241 500,00
24		23 500,00	25 000,00	9 500,00	61 500,00		119 500,00
25	87 000,00	18 500,00		64 500,00	55 500,00	11 500,00	237 000,00
26		34 500,00	9 000,00				43 500,00
27		73 500,00		38 000,00		8 500,00	120 000,00
28	45 500,00	19 500,00	6 500,00	14 500,00	15 000,00		101 000,00
29				13 500,00	61 500,00	15 500,00	90 500,00
30	59 000,00		37 000,00	23 500,00	95 000,00		214 500,00
31					38 500,00		38 500,00
32					10 500,00	45 000,00	55 500,00
33			4 000,00	55 000,00			59 000,00
34			2 500,00	89 500,00	35 000,00		127 000,00

Рис. 7. Транспонированное отображение куба

Таблица X Куб X

Цель кредитования

Возраст

- ☒ Иное
- ☒ Оплата за образование
- ☒ Оплата услуг (мед., юрид. и т.п.)
- ☒ Покупка и ремонт недвижимости
- ☒ Покупка товара
- ☒ Турпоездки, развлечения и т.п.

Возраст	Иное	Оплата за образование	Оплата услуг (мед. и т.п.)	Покупка и ремонт недвижимости	Покупка товара	Турпоездки, развлечения и т.п.	Итого:
19	50 000,00			78 000,00	46 500,00		174 500,00
20	17 000,00	17 500,00			73 500,00	30 500,00	138 500,00
21	8 500,00	29 500,00		13 000,00	76 500,00		127 500,00
22	23 500,00	31 500,00		46 500,00	112 000,00		213 500,00
23		49 000,00	66 500,00	27 500,00	98 500,00		241 500,00
24		23 500,00	25 000,00	9 500,00	61 500,00		119 500,00
25	87 000,00	18 500,00		64 500,00	55 500,00	11 500,00	237 000,00
26		34 500,00	9 000,00				43 500,00
27		73 500,00		38 000,00		8 500,00	120 000,00
28	45 500,00	19 500,00	6 500,00	14 500,00	15 000,00		101 000,00
29				13 500,00	61 500,00	15 500,00	90 500,00
30	59 000,00		37 000,00	23 500,00	95 000,00		214 500,00
31					38 500,00		38 500,00
32					10 500,00	45 000,00	55 500,00
33			4 000,00	55 000,00			59 000,00
34			2 500,00	89 500,00	35 000,00		127 000,00
35			20 500,00	19 000,00			39 500,00
36			17 000,00	26 500,00	45 500,00		89 000,00

Выделение по маске

Содержит: Оплата

☒ Отметить элементы
☐ Добавить к выбранным
☐ Снять выделение

☐ С учетом регистра
☒ Без учета регистра

Ok Отмена

Рис. 8. Задание параметров фильтрации с помощью инструмента Выделение по маске

14. Выполнить эксперименты с использованием в режиме представления Куба других инструментальных кнопок (Селектор..., Показывать итоги, Настройка фактов и др.).

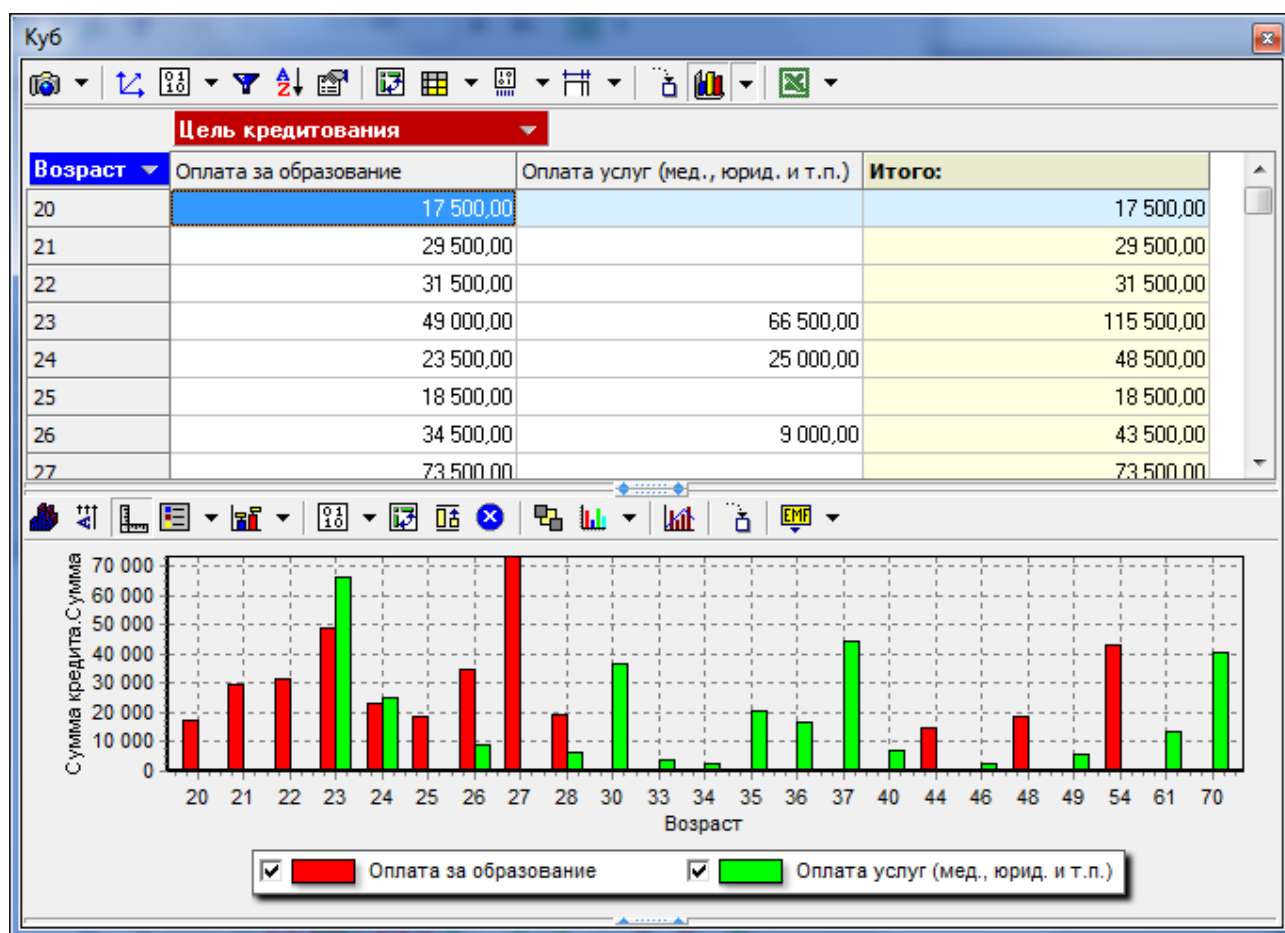


Рис. 9. Результат фильтрации куба в виде таблицы

15. Используя кнопку области диаграммы *Параметры тренда*, задать представление тренда в виде вейвлет-преобразования (рис. 10).

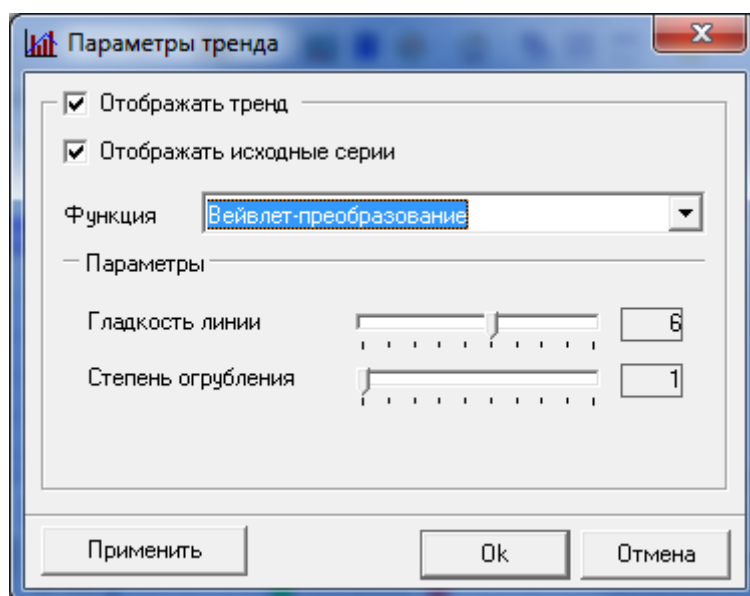


Рис. 10. Окно задания параметров тренда

16. Исследовать и объяснить результаты операции задания тренда (рис. 11).
17. Выполнить сохранение конфигурации с помощью кнопки *Управление конфигурациями*.
18. Выполнить эксперименты с другими параметрами тренда.

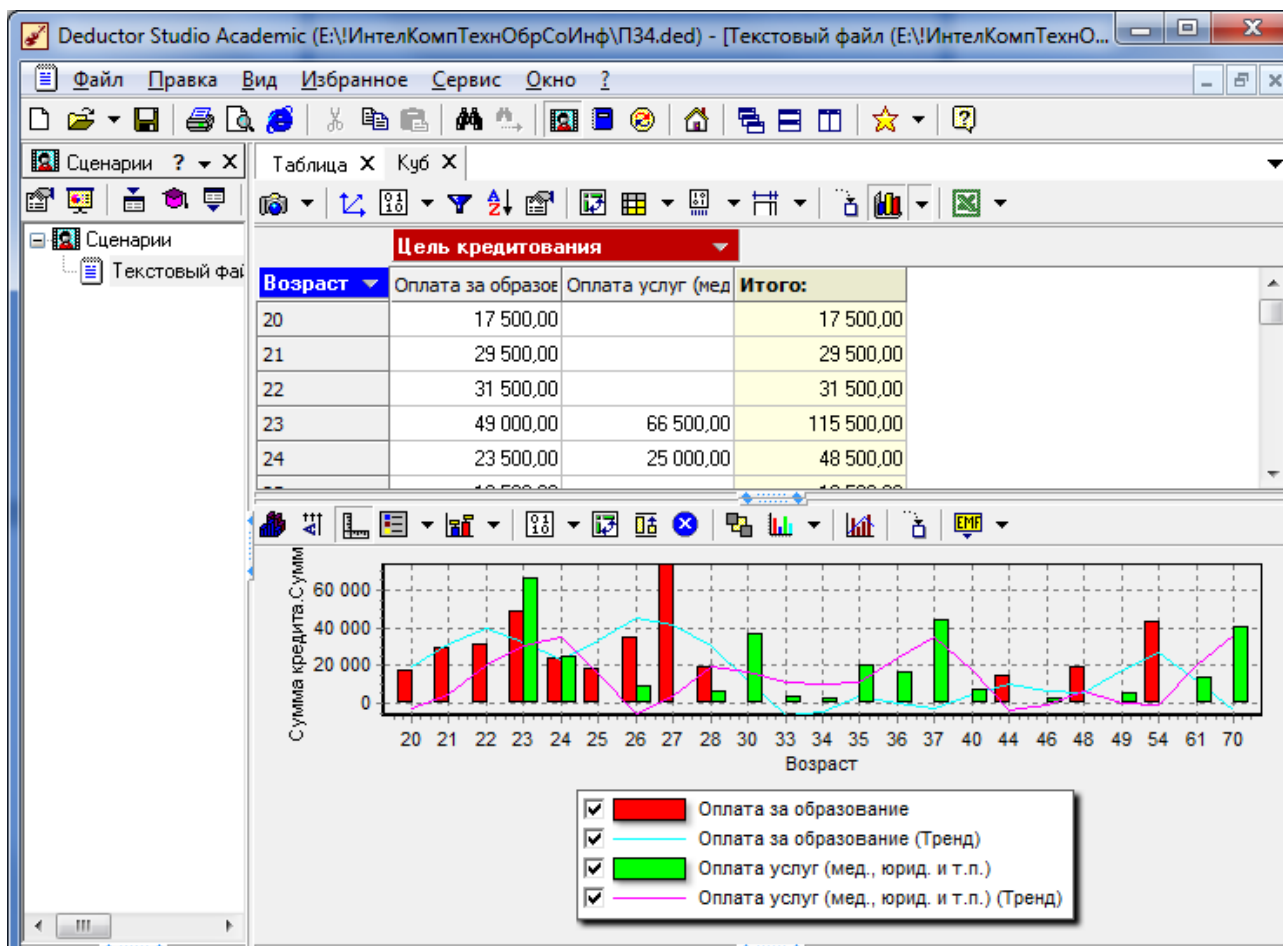


Рис. 11. Результаты обработки исходных данных с использованием многомерного представления (куба), фильтрации и линии тренда в виде вейвлет-преобразования

19. Задать другие параметры отбора значений измерения *Цель кредитования* (например, по столбцам *Покупка товаров*, *Покупка и ремонт недвижимости*). Представить диаграмму в 3-х мерном виде (рис. 12). Выполнить анализ.
20. Сохранить результат под именем *Конфигурация № 2*.
21. Выполнить переход от одной к другой конфигурации (от Конфигурации №2 к Конфигурации №1). Проанализировать результаты.
22. Перейти в режим представления *Таблица*. Включить отображение *статистики* (рис. 13). Выполнить анализ.
23. Задать параметры фильтрации табличных данных: выдать информацию о кредитах женщин в возрасте от 25 до 30 лет включительно на оплату образования, а также мужчин не старше 22 лет, имеющих импортный автомобиль (рис. 14).
24. Проанализировать результаты (рис. 15). Сохранить результаты как Конфигурацию №3.
25. Поэкспериментировать с заданием условий фильтрации.

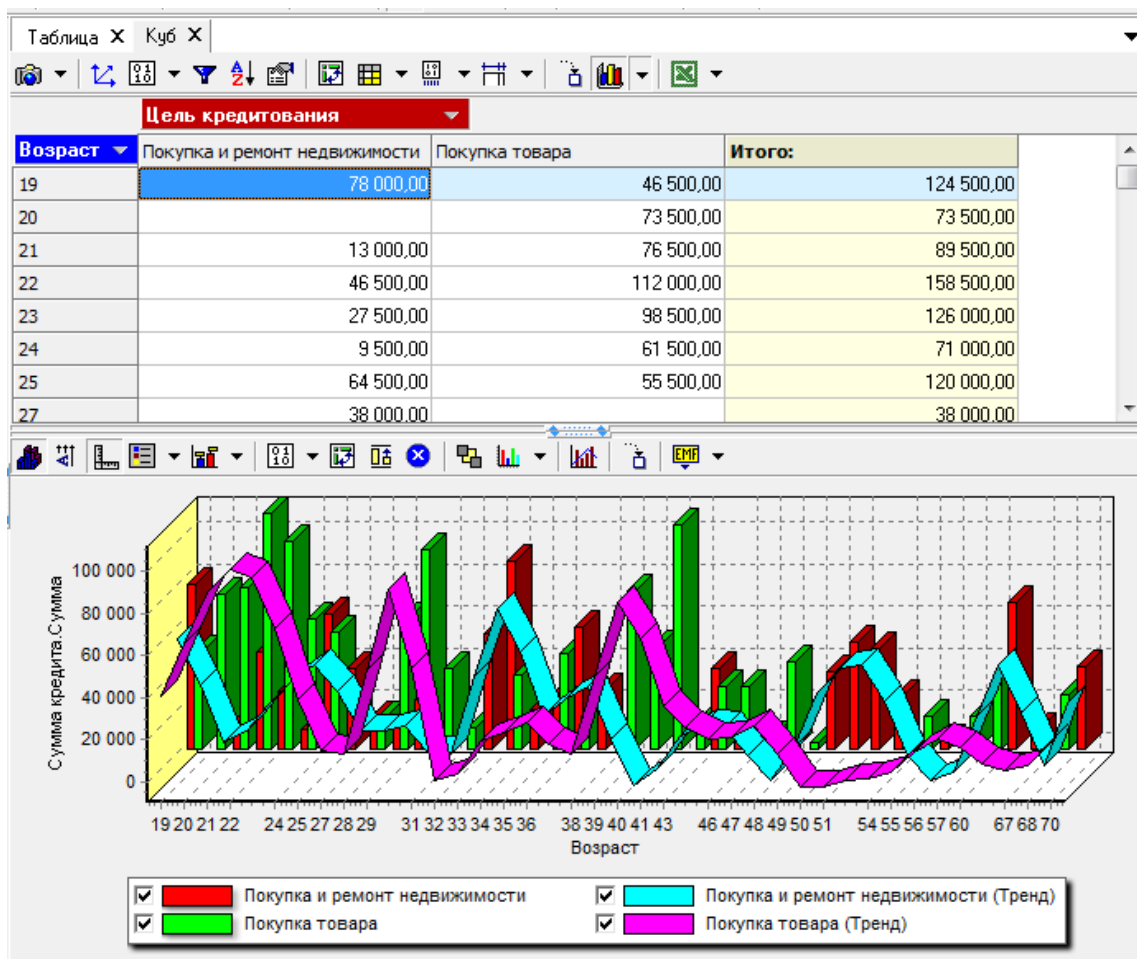


Рис. 12. Представление результатов для Конфигурации №2

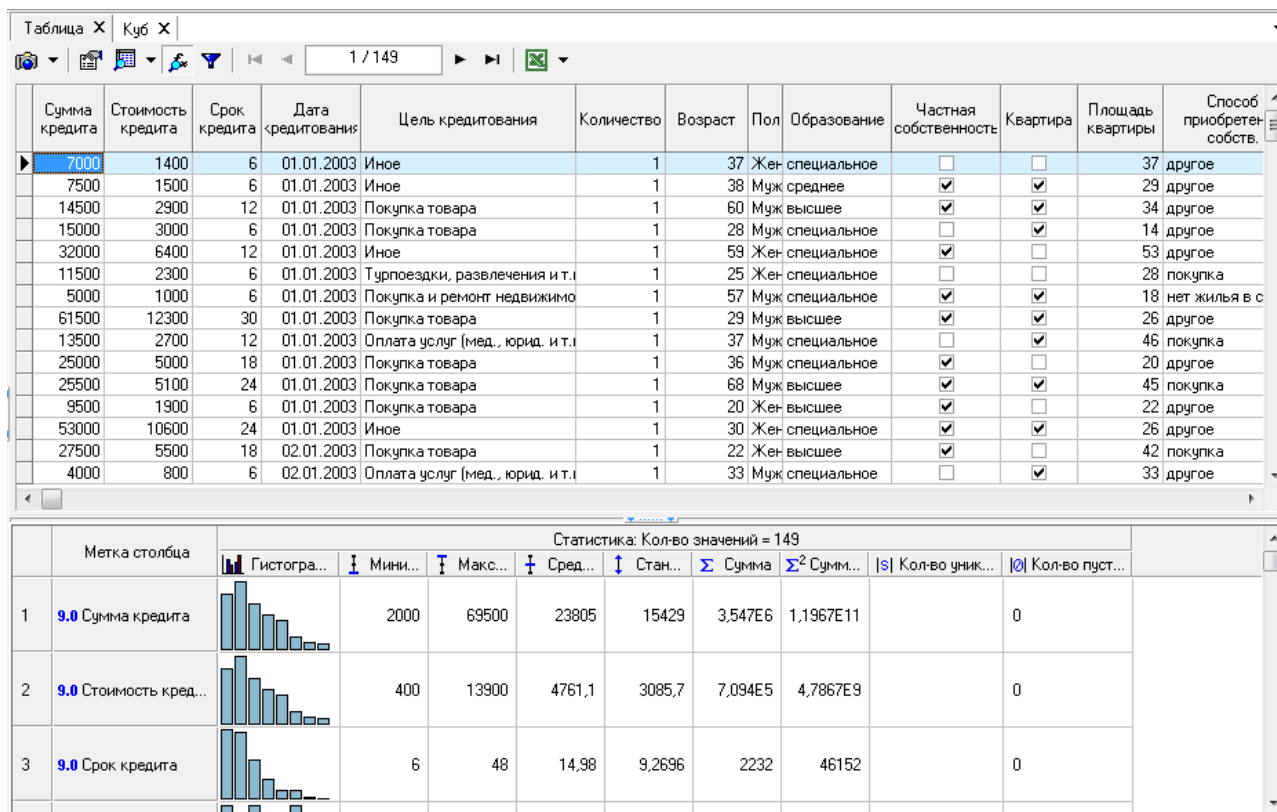


Рис. 13. Отображение статистики

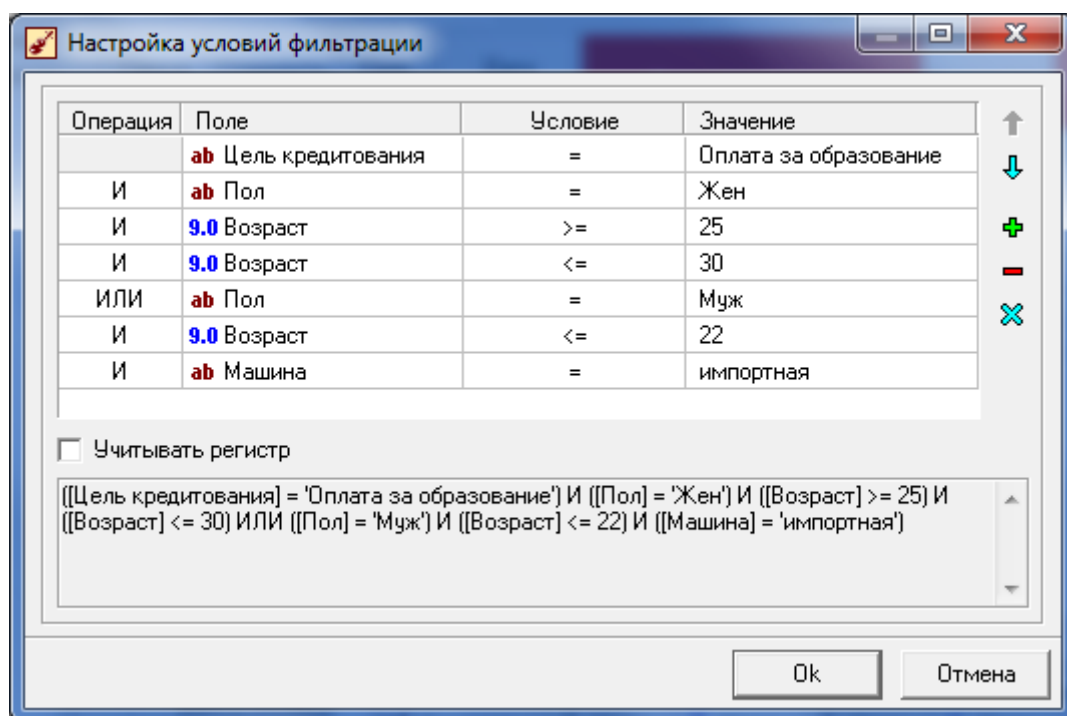


Рис. 14. Задание параметров фильтрации табличных данных

Таблица

1 / 7

Сумма кредита	Стоимость кредита	Срок кредита	Дата кредитования	Цель кредитования	Количество	Возраст	Пол	Образование	Частная собственность	Квартира	Площадь
23500	4700	12	05.01.2003	Оплата за образо	1	27	Жен	специальное	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
31000	6200	18	06.01.2003	Оплата за образо	1	27	Жен	специальное	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
12000	2400	6	07.01.2003	Покупка и ремонт	1	22	Муж	специальное	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
6000	1200	6	08.01.2003	Покупка товара	1	22	Муж	специальное	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
19000	3800	18	08.01.2003	Оплата за образо	1	27	Жен	специальное	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
31500	6300	24	09.01.2003	Оплата за образо	1	22	Муж	специальное	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
29500	5900	18	11.01.2003	Оплата за образо	1	21	Муж	высшее	<input type="checkbox"/>	<input checked="" type="checkbox"/>	

Статистика: Кол-во значений = 7

Метка столбца	Гистогра...	Мини...	Макс...	Сред...	Стан...	Σ Сумма	Σ² Сумм...	s Кол-во уника...	0 K
1 9.0 Сумма кредита		6000	31500	21786	9953,5	1,525E5	3,9168E9		0
2 9.0 Стоимость кред...		1200	6300	4357,1	1990,7	30500	1,5667E8		0

Рис. 15. Результат фильтрации табличных данных

26. Перейти в режим представления результатов в виде куба. Задать условия фильтрации с помощью кнопки *Селектор* (для фактов – рис. 16, для цели кредитования – рис. 17, для возраста – рис. 18).
27. Результат фильтрации (рис. 19) сохранить в виде конфигурации № 4. Сохранить проект на диске в личной папке.

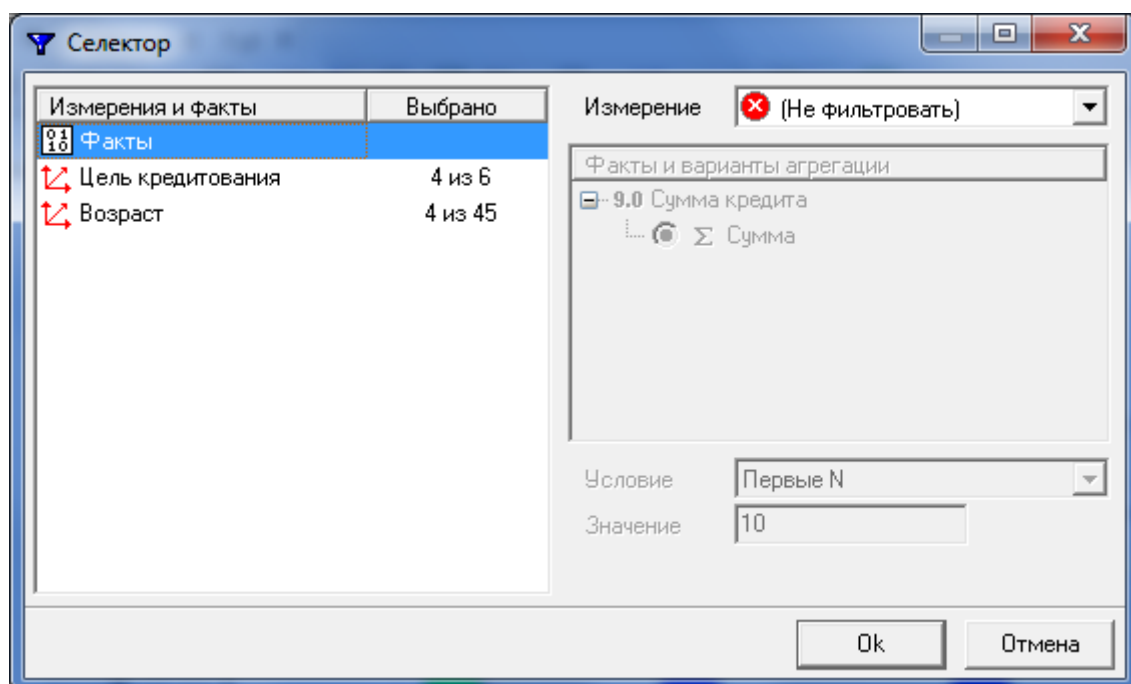


Рис. 16. Окно задания условий фильтрации для фактов

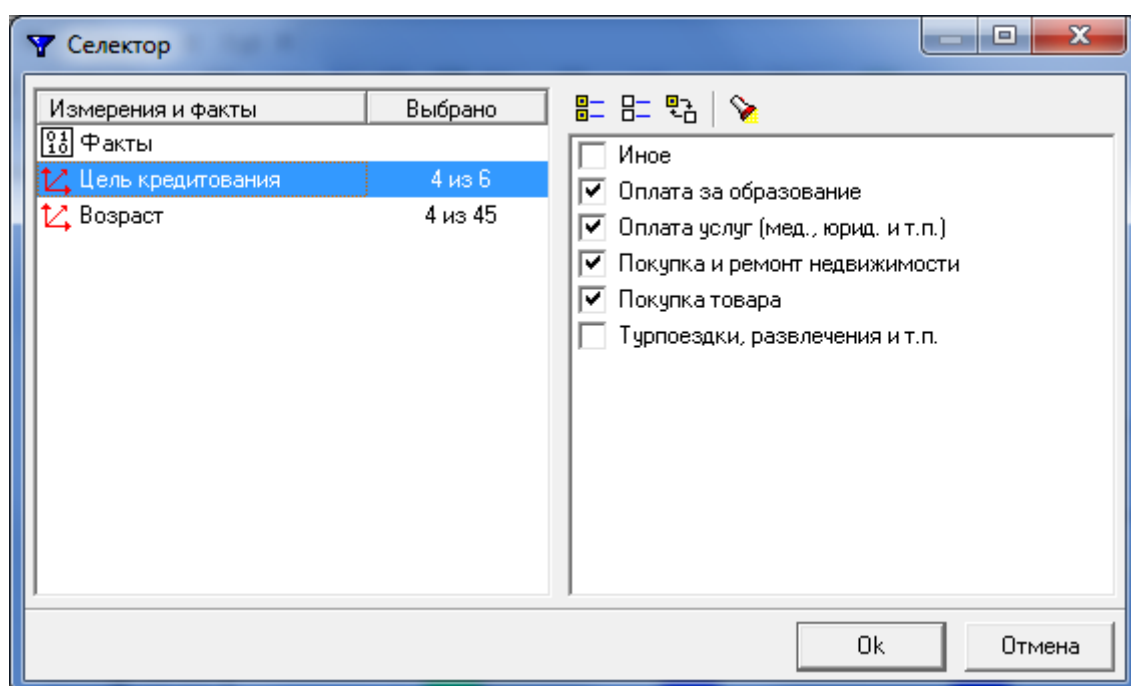


Рис. 17. Окно задания условий фильтрации для цели кредитования

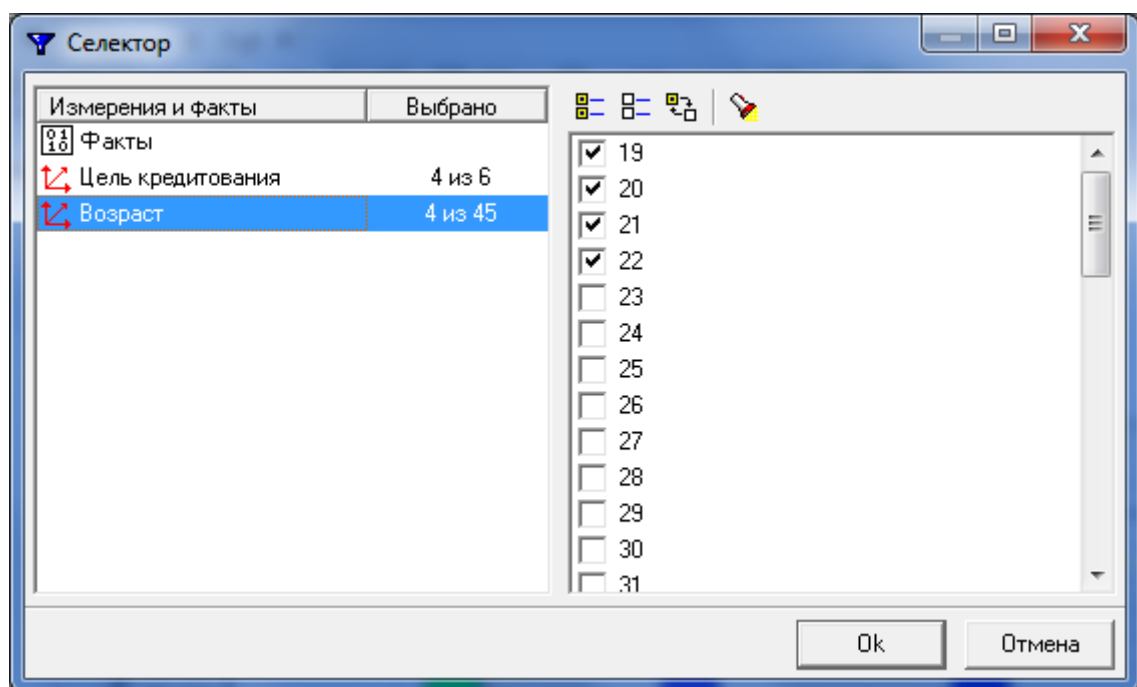


Рис. 18. Окно задания условий фильтрации для возраста

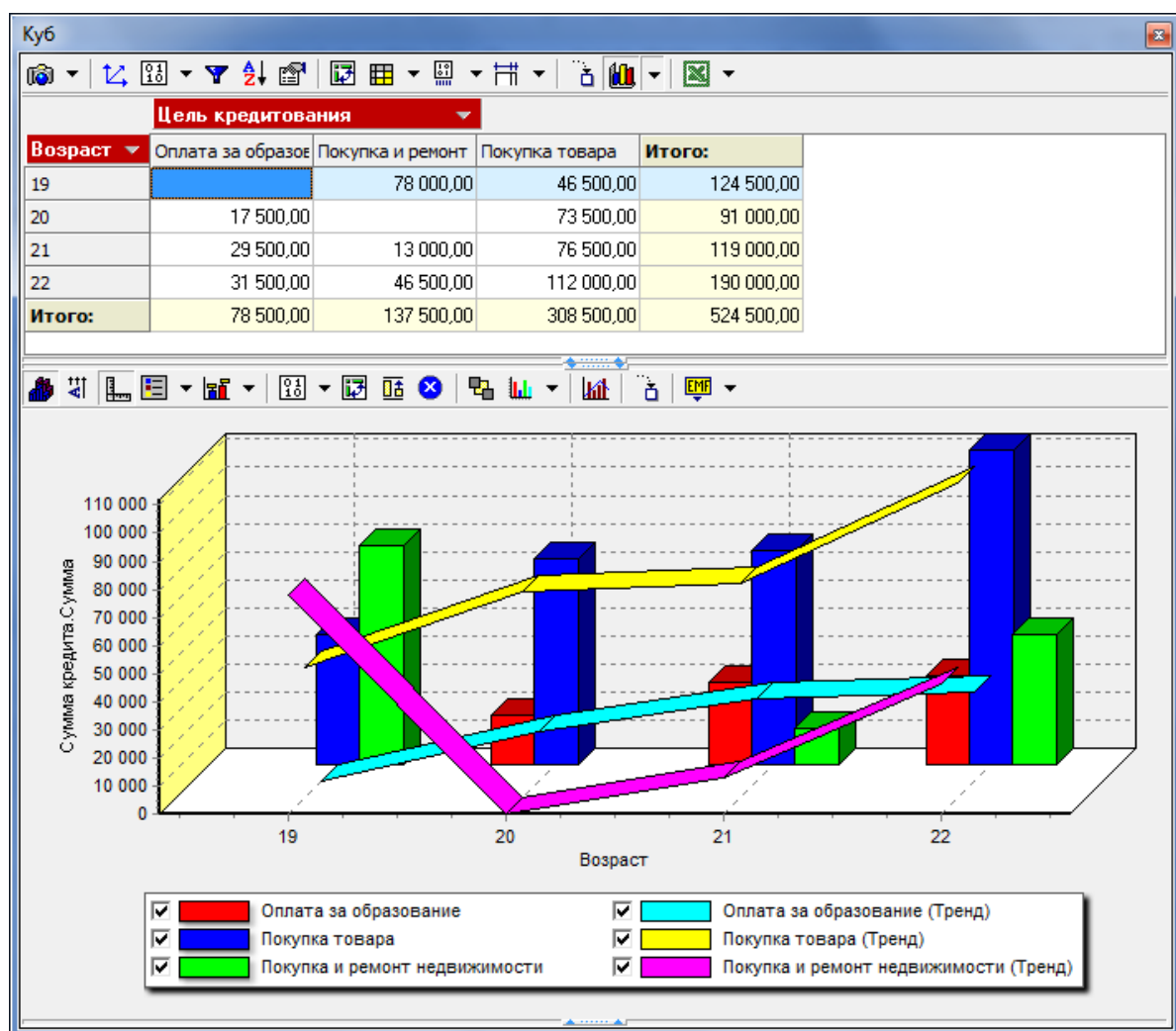


Рис. 19. Результаты фильтрации

2. Разбиение дат по временным отрезкам

Разбиение дат по временным отрезкам служит для анализа всевозможных показателей за определенный период (день, неделя, месяц, квартал, год). Суть разбиения заключается в том, что на основе столбца с информацией о дате формируется другой столбец, в котором указывается, к какому заданному интервалу времени принадлежит строка данных. Тип интервала задается аналитиком, исходя из того, что он хочет получить – данные за год, квартал, месяц, неделю, день или сразу по всем интервалам.

Исходные данные: текстовый файл **Credit.txt**.

Требуется: получить сведения по суммам взятых кредитов по неделям (в файле Credit.txt содержится информация за первые две недели 2003 года).

1. Запустить Мастер обработки. В группе Трансформация данных выбрать для преобразования даты и времени метод обработки Дата и время (рис. 20). Перейти к следующему шагу.

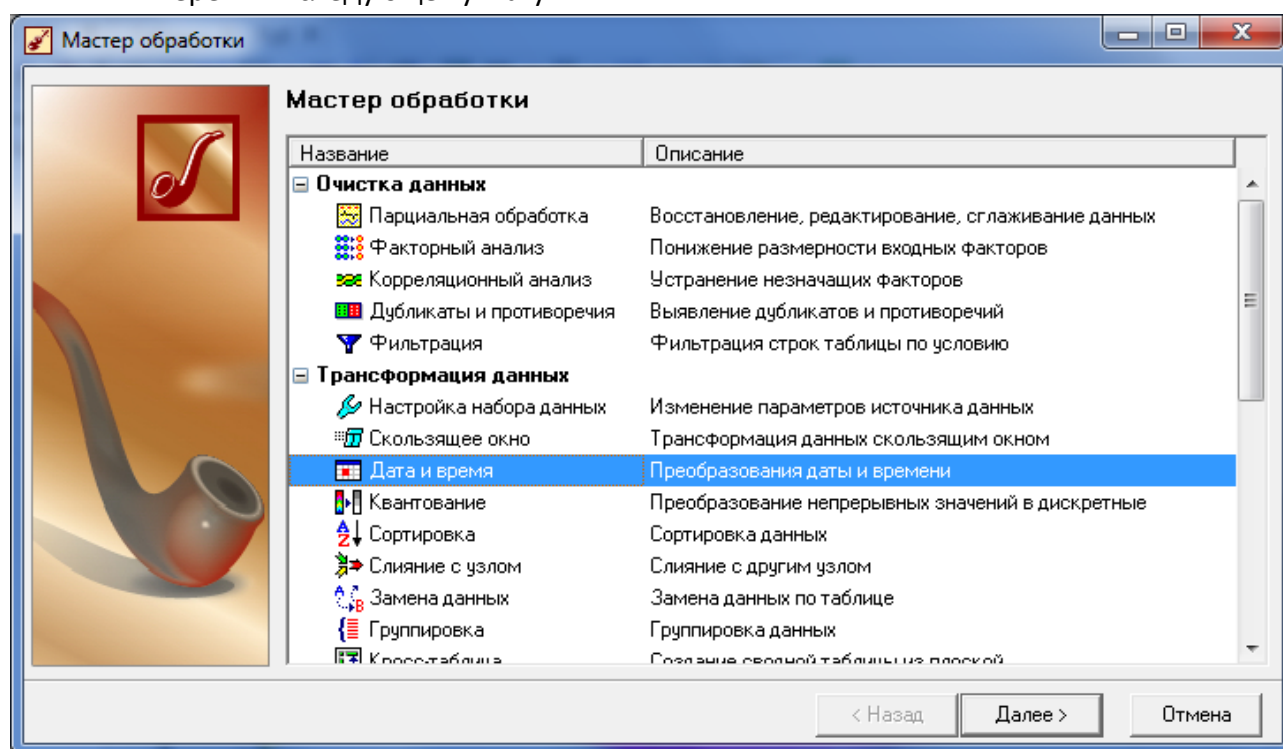


Рис. 20. Окно выбора метода обработки данных

2. На шаге 2 Мастера обработки для столбца *Дата кредитования* установить флажок *Год + Неделя* (рис. 21).
3. На шаге 3 Мастера обработки установить флажки отображения результатов в виде таблицы и куба. Перейти к следующему шагу.
4. На шаге 4 Мастера обработки в качестве измерений установить поля *Год+Неделя* и *Цель кредитования*, а в качестве факта – поле *Сумма кредита*. Остальные поля сделать неиспользуемыми (рис. 22).
5. На шаге 5 Мастера обработки разместить поля измерений: *Год+Неделя* – в колонках, *Цель кредитования* – в строках (рис. 23).

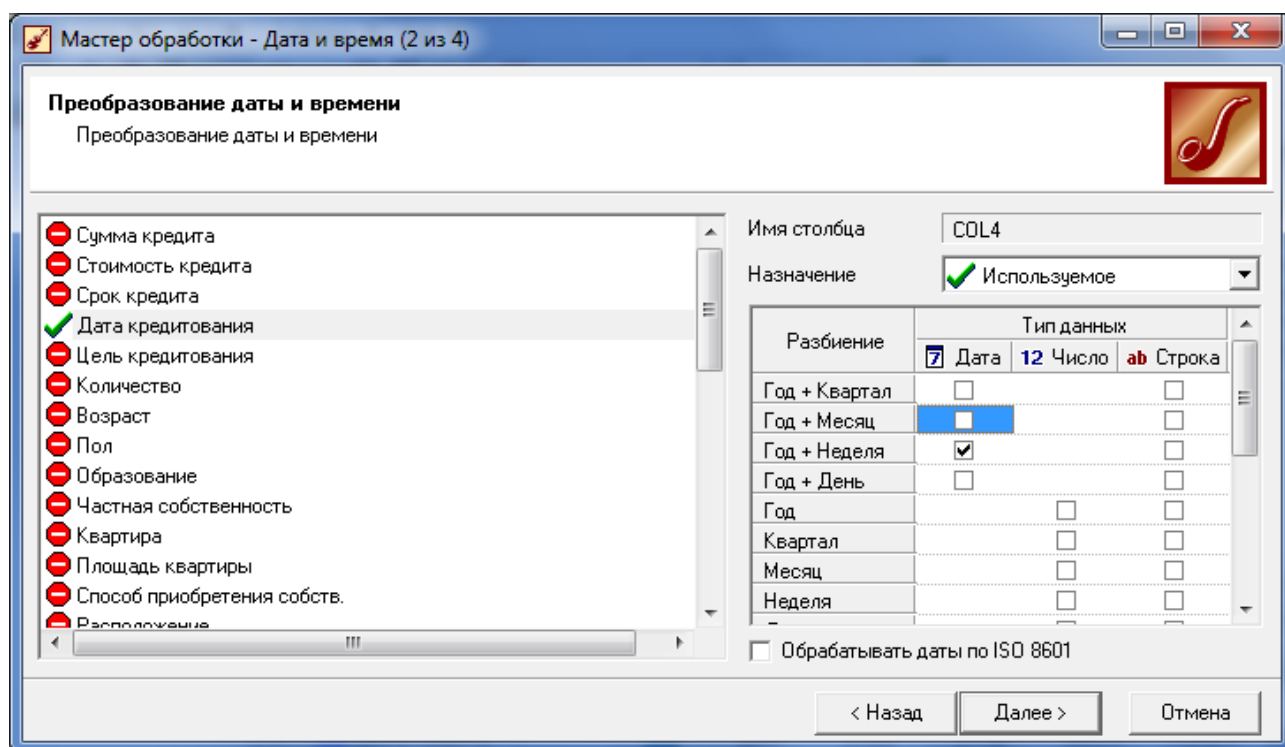


Рис. 21. Задание алгоритма преобразования поля Дата кредитования

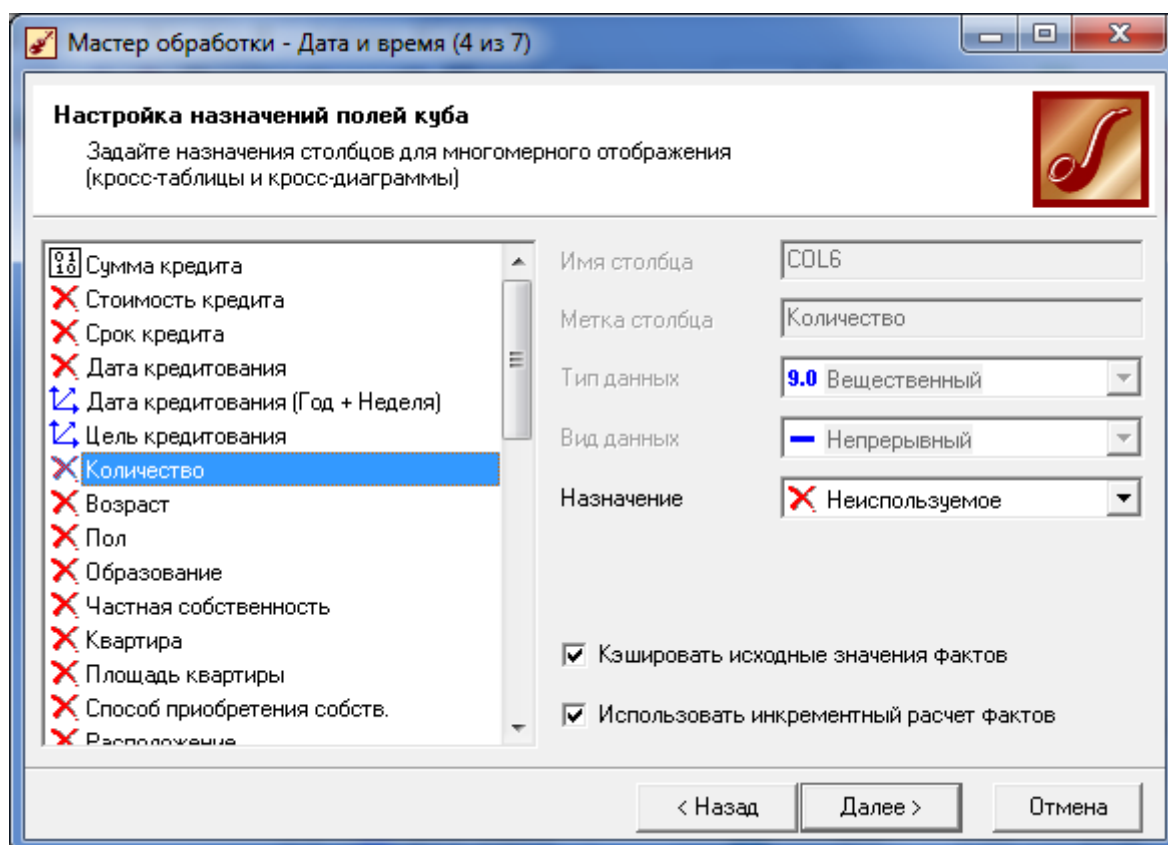


Рис. 22. Настройка назначения полей куба

6. На шаге 6 Мастера обработки выполнить настройку факта – отображать сумму кредитов и их количество (рис. 24).

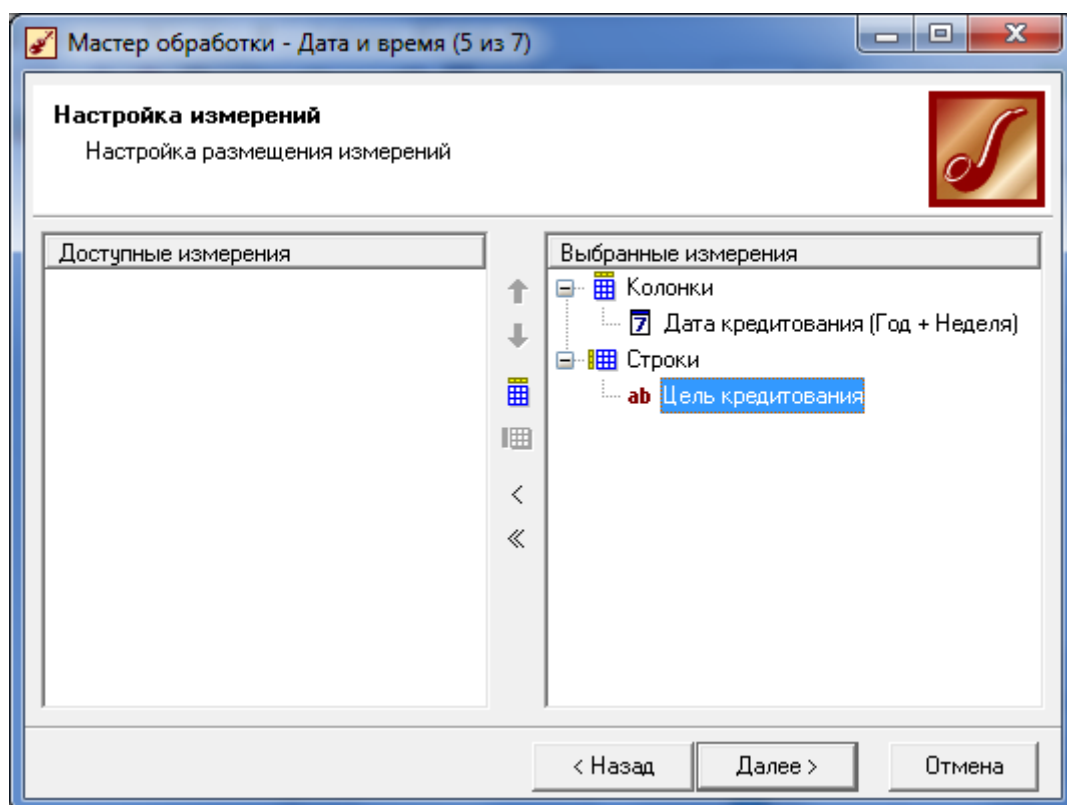


Рис. 23. Настройка размещения измерений

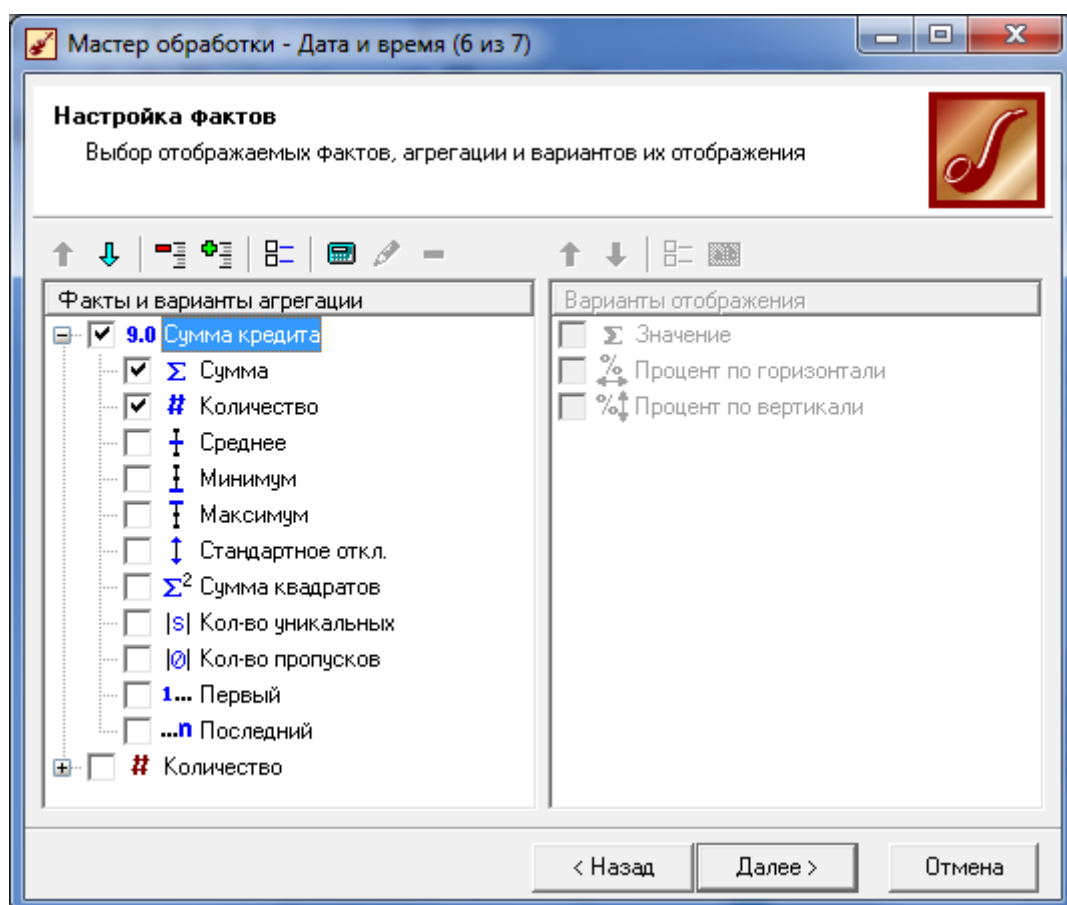


Рис. 24. Настройка фактов, агрегации и вариантов их отображения

7. Перейти к завершению обработки. Результат в виде куба представлен на рис. 25.

8. Провести анализ полученных данных. (Отметим, что первая неделя в 2003 году была неполной, а вторая неделя началась 6 января).
9. Сохранить результаты в виде новой конфигурации.

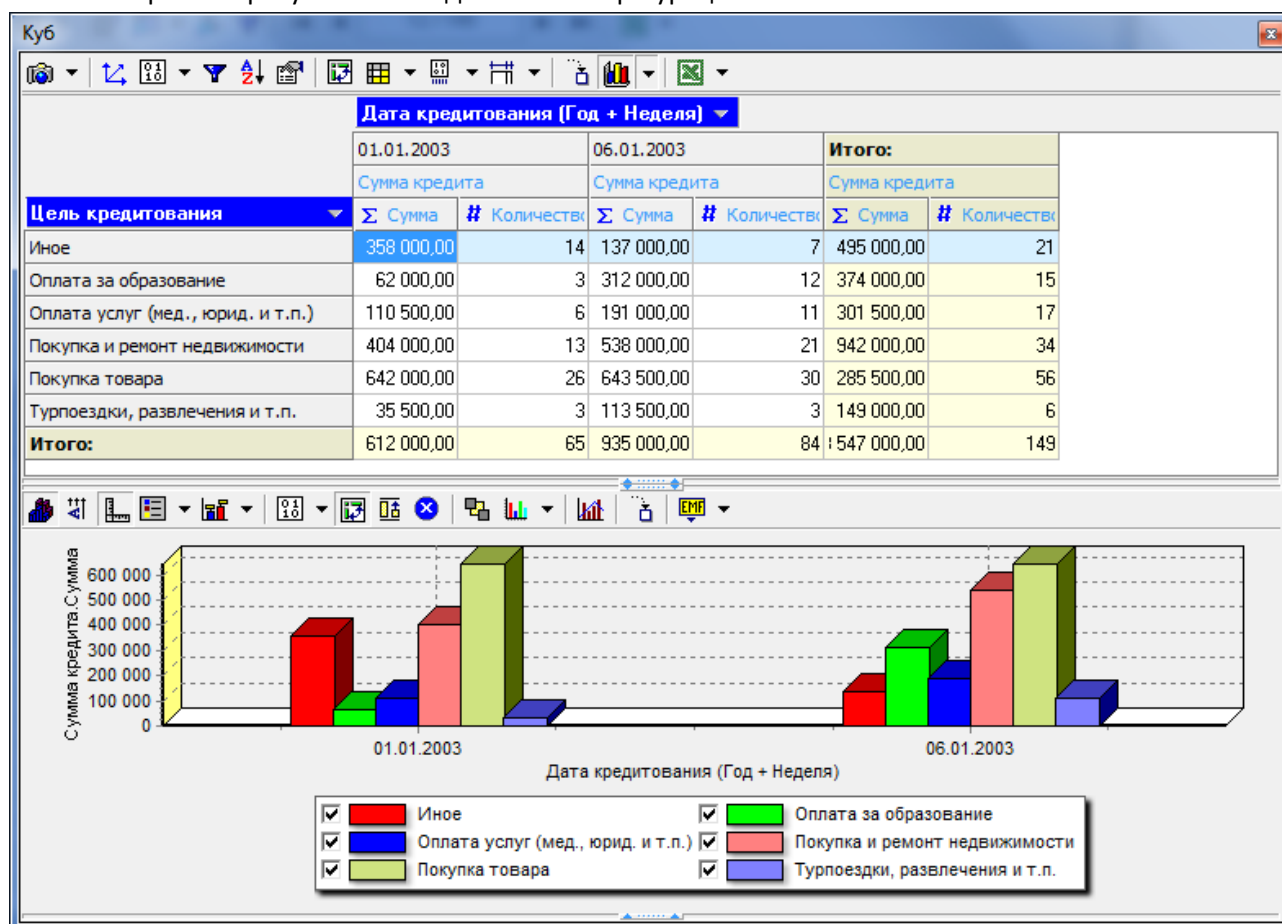


Рис. 25. Результат отображения данных по выданным кредитам по неделям

3. Квантование возраста заемщиков по интервалам

Часто аналитику необходимо отнести непрерывные данные (например, количество продаж) к какому-либо конечному набору. Например, всю совокупность данных о количестве продаж необходимо разбить на 5 интервалов – от 0 до 100, от 100 до 200 и т.д., и отнести каждую запись исходного набора к какому-то конкретному интервалу и далее проводить анализ или фильтрации, исходя именно из этих интервалов.

Для выполнения указанной операции в Deductor Studio применяется инструмент квантования (или дискретизации).

Квантование предназначено для преобразования непрерывных данных в дискретные.

Преобразование может проходить как по *интервалам* (данные разбиваются на заданное количество интервалов одинаковой длины), так и по *квантилям* (данные разбиваются на интервалы разной длины так, чтобы в каждом интервале находилось одинаковое количество данных). В качестве значений результирующего набора данных могут выступать номер интервала, нижняя или верхняя граница интервала, середина интервала, либо метка интервала (значения определяемые аналитиком).

Примером использования квантования может служить разбиение данных о возрасте заемщиков на пять интервальных групп:

- до 30 лет,
- от 30 до 40 лет,
- от 40 до 50 лет,
- от 50 до 60 лет,
- старше 60 лет.

Исходные данные распределятся по пяти интервалам именно так, поскольку, согласно статистике, минимальное значение возраста заемщика 19, а максимальное 69 лет.

Квантование позволит аналитику оценить кредиторскую активность представителей разных возрастных групп с целью принятия решения о стимулировании заемщиков в группах с низкой активностью (например, путем уменьшения стоимости кредита для представителей этих групп) и, быть может, увеличения прибыли в возрастных группах заемщиков с высоким риском (путем увеличения для них стоимости кредита).

Поскольку аналитика могут интересовать данные в разрезе по неделям, поэтому продолжим работу на основе последних полученных результатов обработки текстового файла **Credit.txt**.

1. Запустить Мастер обработки. В группе Трансформация данных выбрать метод обработки Квантование. Перейдем к следующему шагу.
2. На шаге 2 Мастера обработки задать параметры квантования для поля Возраст: Назначение – Используемое, Способ – По интервалам, Интервалов – 5, Значение – Метка интервала, Вид данных – Дискретный (рис. 26).

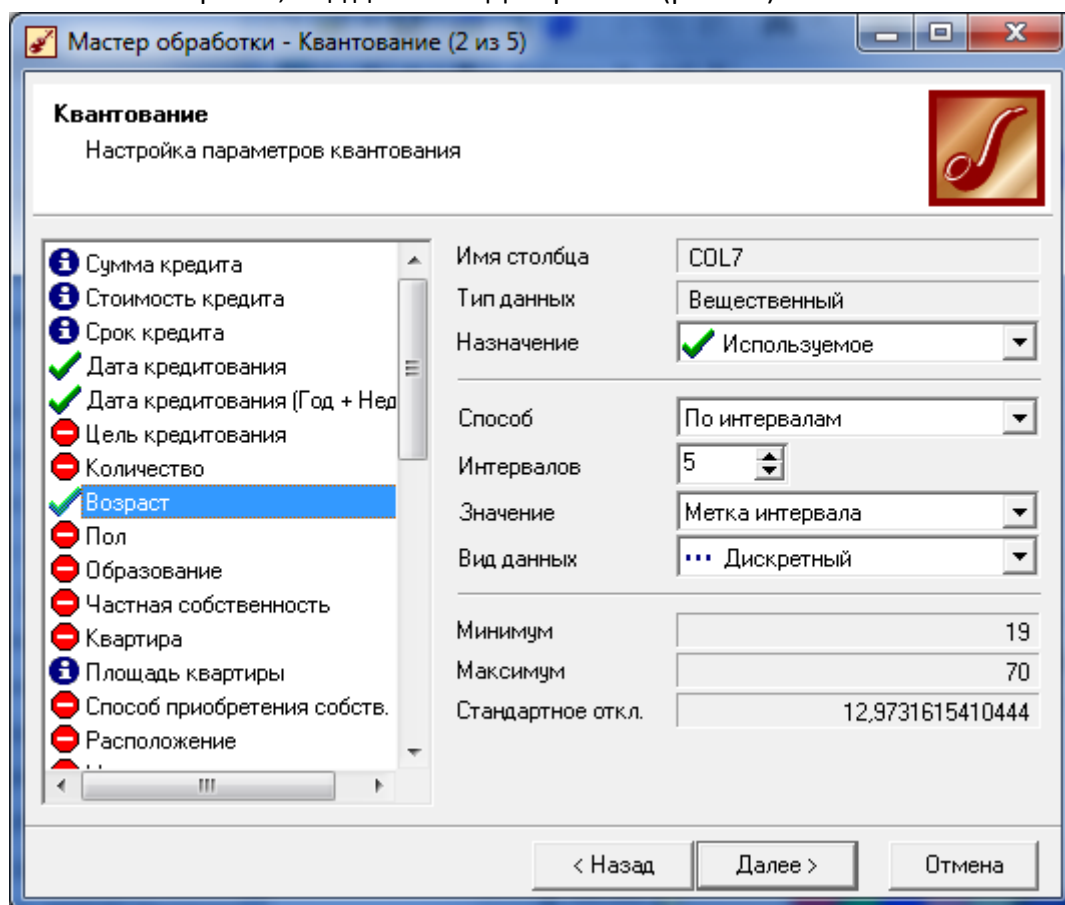


Рис. 26. Задание квантования по интервалам для возраста заемщиков

3. На шаге 3 Мастера обработки задать метки возрастным интервалам: *До 30 лет, От 30 до 40 лет, От 40 до 50 лет, От 50 до 60 лет, Свыше 60 лет* (рис. 27).

Мастер обработки - Квантование (3 из 5)

Границы и метки интервалов
Настройка границ и меток интервалов квантования

Столбцы	
Имя	Интервалов
7 Дата кредитования	8
7 Дата кредитования (Г...	8
9.0 Возраст	5

Интервалы		
№	Граница	Метка
	19	
0	29,2	До 30 лет
1	39,4	От 30 до 40 лет
2	49,6	от 40 до 50 лет
3	59,8	От 50 до 60 лет
4	70	Свыше 60 лет

< Назад Далее > Отмена

Рис. 27. Задание меток возрастным интервалам

4. На шаге 4 Мастера обработки задать представление в виде таблицы и куба.
5. На шаге 5 Мастера обработки указать в качестве измерений поля *Возраст* и *Год+Неделя*, в качестве факта – *Сумма кредита*. Остальные поля указать, как неиспользуемые (рис. 28).
6. На шаге 6 Мастера обработки разместить поля измерений: *Год+Неделя* – в колонках, *Возраст* – в строках (рис. 29).
7. На шаге 7 Мастера обработки задать параметры настройки факта (рис. 30).
8. Перейти к завершению работы Мастера обработки.
9. Отобразить результаты в виде куба. При необходимости выполнить сортировку результирующей таблицы. Отобразить ниже табличного представления куба диаграмму (рис. 31). Провести анализ полученных результатов.
10. Сохранить конфигурацию.
11. Запустить Мастер визуализации. На шаге 5 Мастера визуализации задать новые параметры факта: Факты и варианты агрегации – *Сумма* и *Количество*, Варианты отображения – *Значение*, *Процент по горизонтали*, *Процент по вертикали* (рис. 32).
12. Перейти к завершению работы Мастера визуализации. Отобразить результаты в виде куба с представлением таблицы и столбчатой диаграммы. При необходимости настроить вид диаграммы (рис. 33) – задать метки и легенду.
13. Изменить вид диаграммы на круговую диаграмму. Результаты представлены на рис. 34. Обратите внимание, что информация подается за весь период кредитования (проценты по возрастным группам соответствуют значениям суммы кредита в области *Итого*).
14. Выполнить сохранение текущей конфигурации под новым именем.

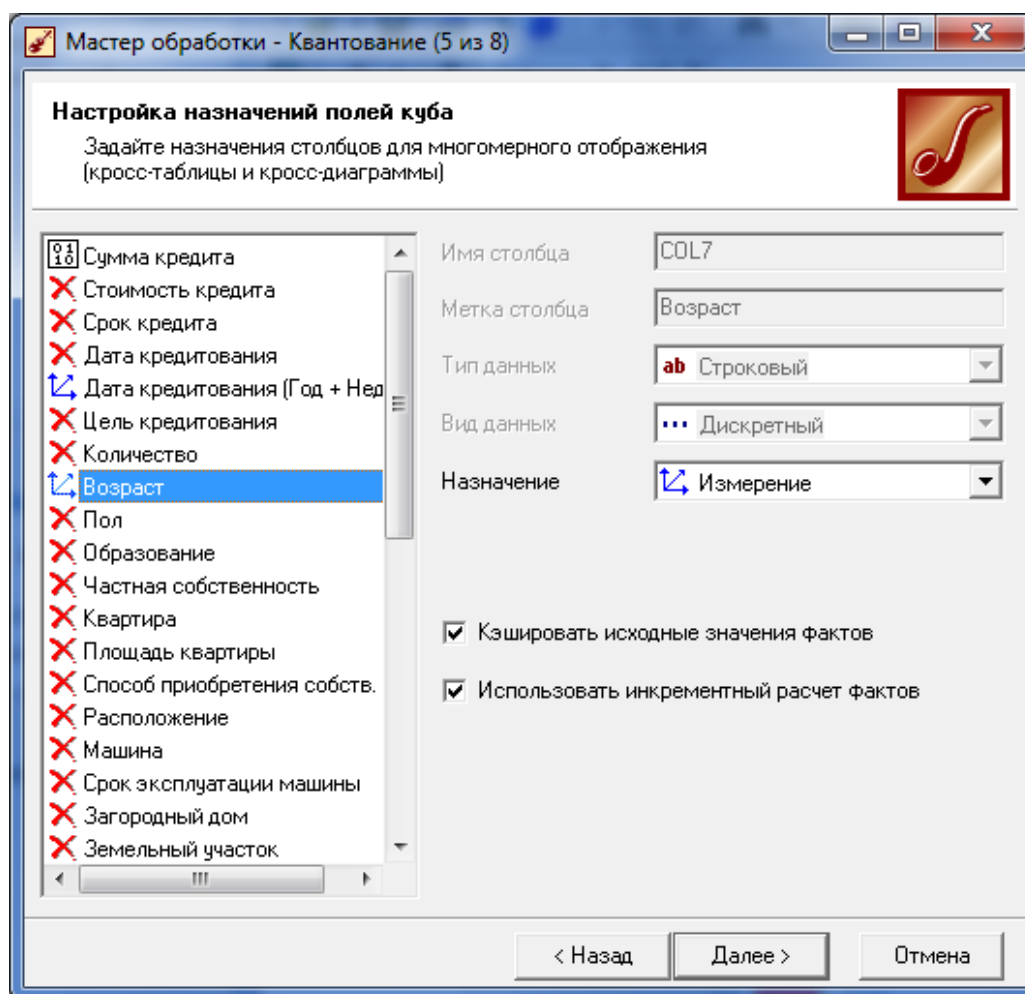


Рис. 28. Настройка назначений полей куба при квантовании по возрасту

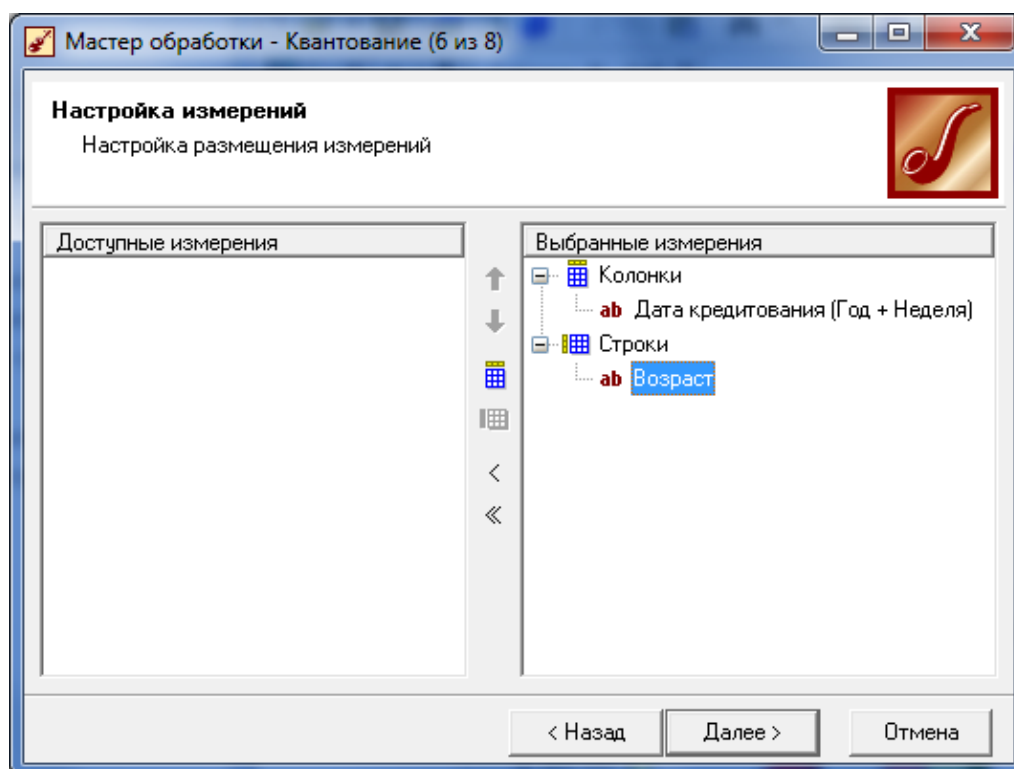


Рис. 29. Настройка измерений при квантовании по возрасту

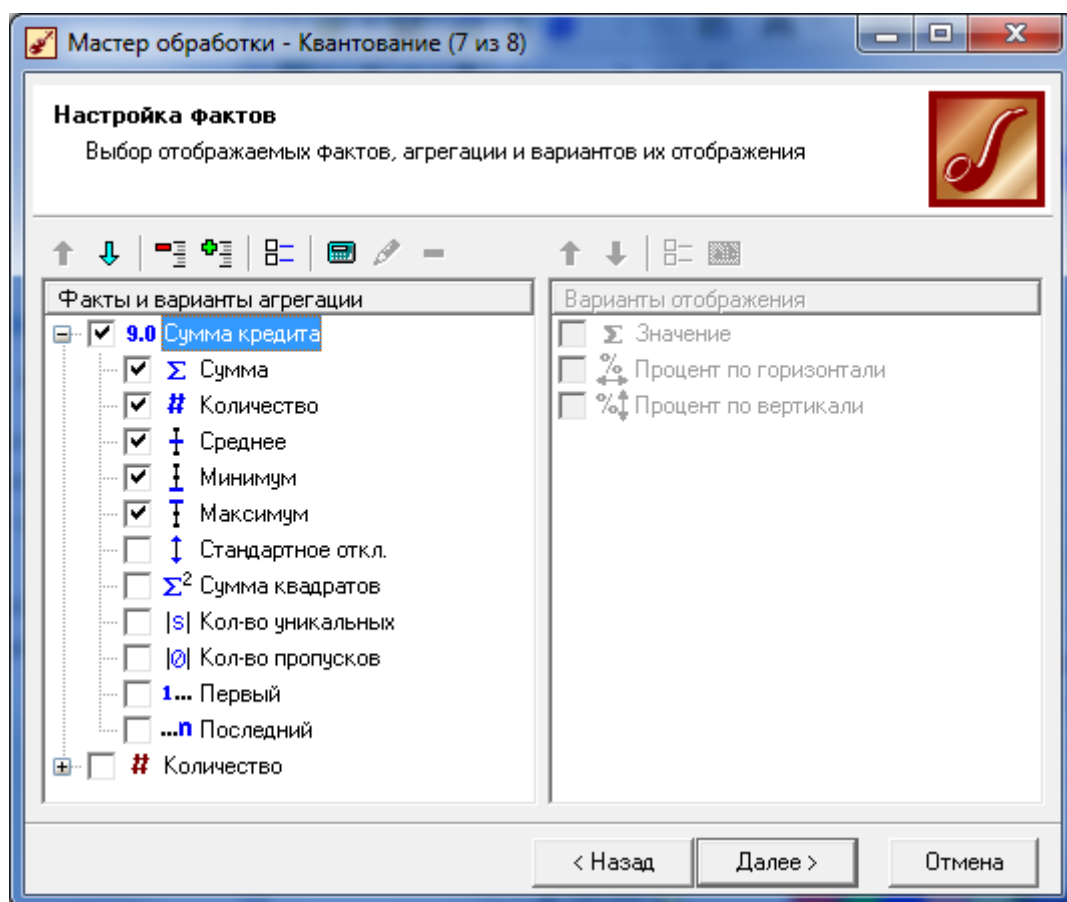


Рис. 30, Параметры настройки фактов при квантовании по возрасту

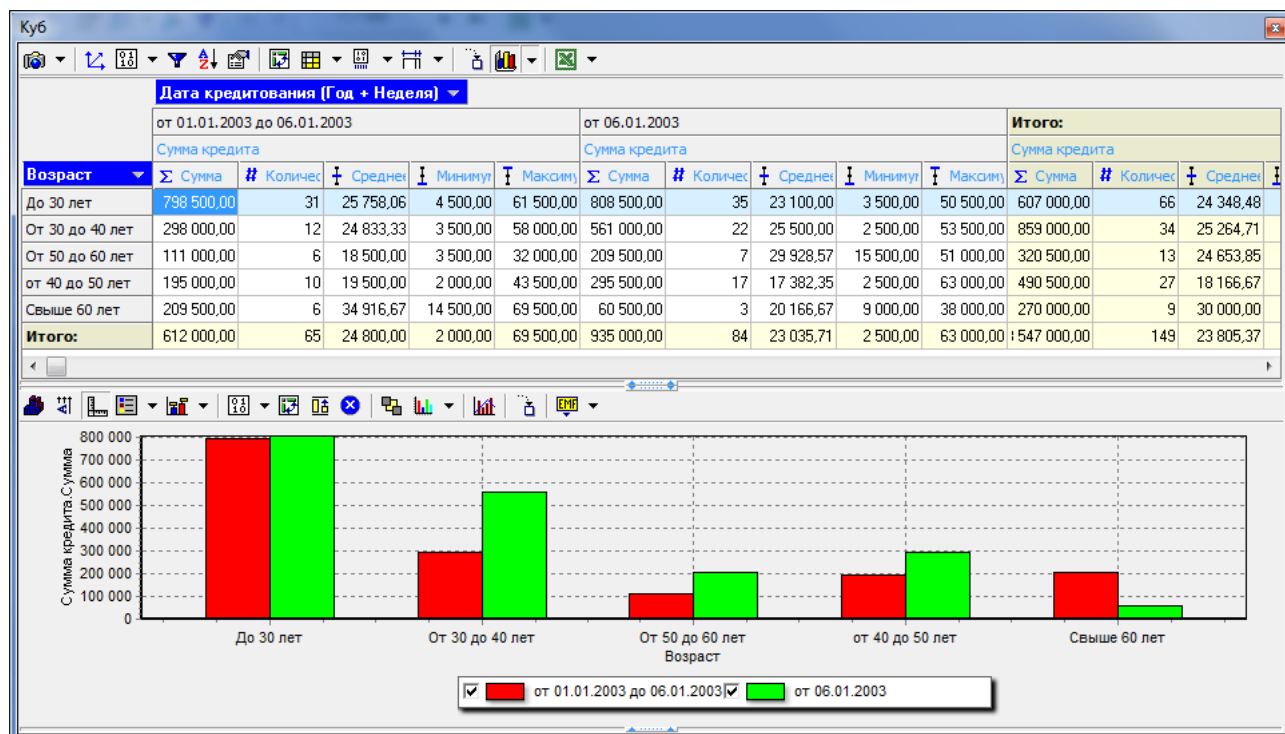


Рис. 31. Результат квантования по возрасту и неделям

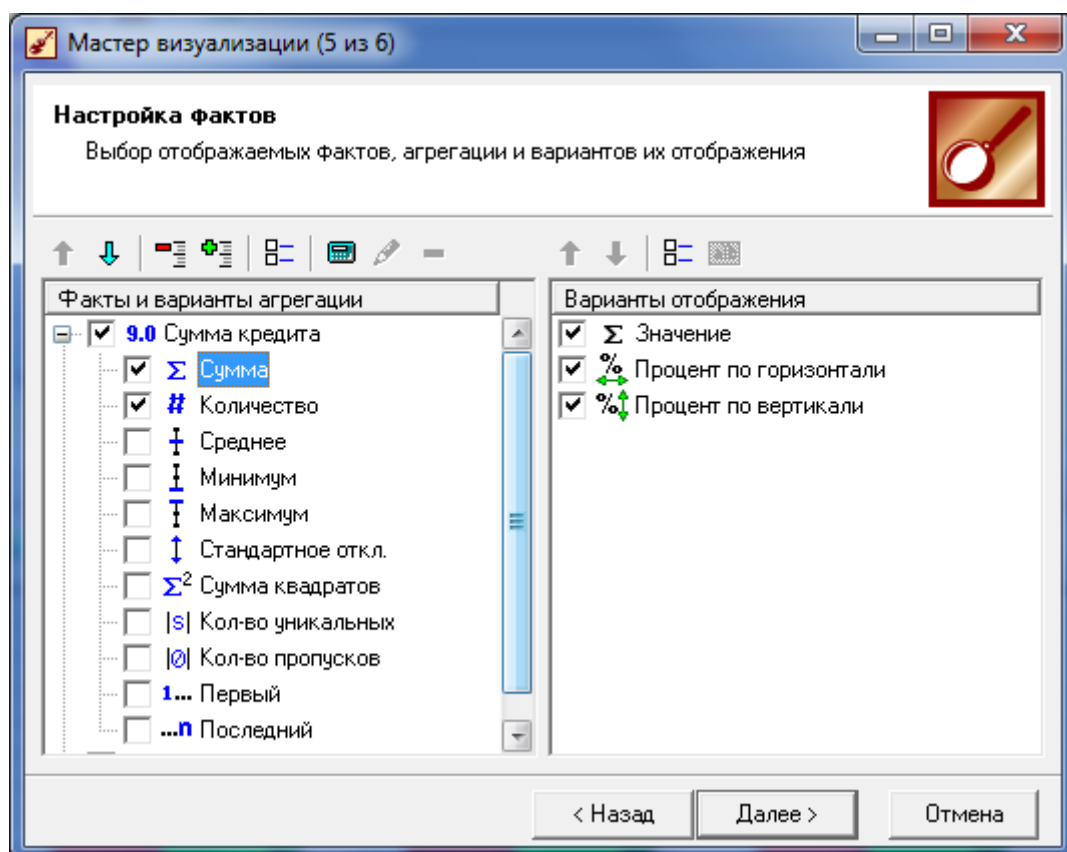


Рис. 32. Новые параметры настройки фактов с помощью Мастера визуализации

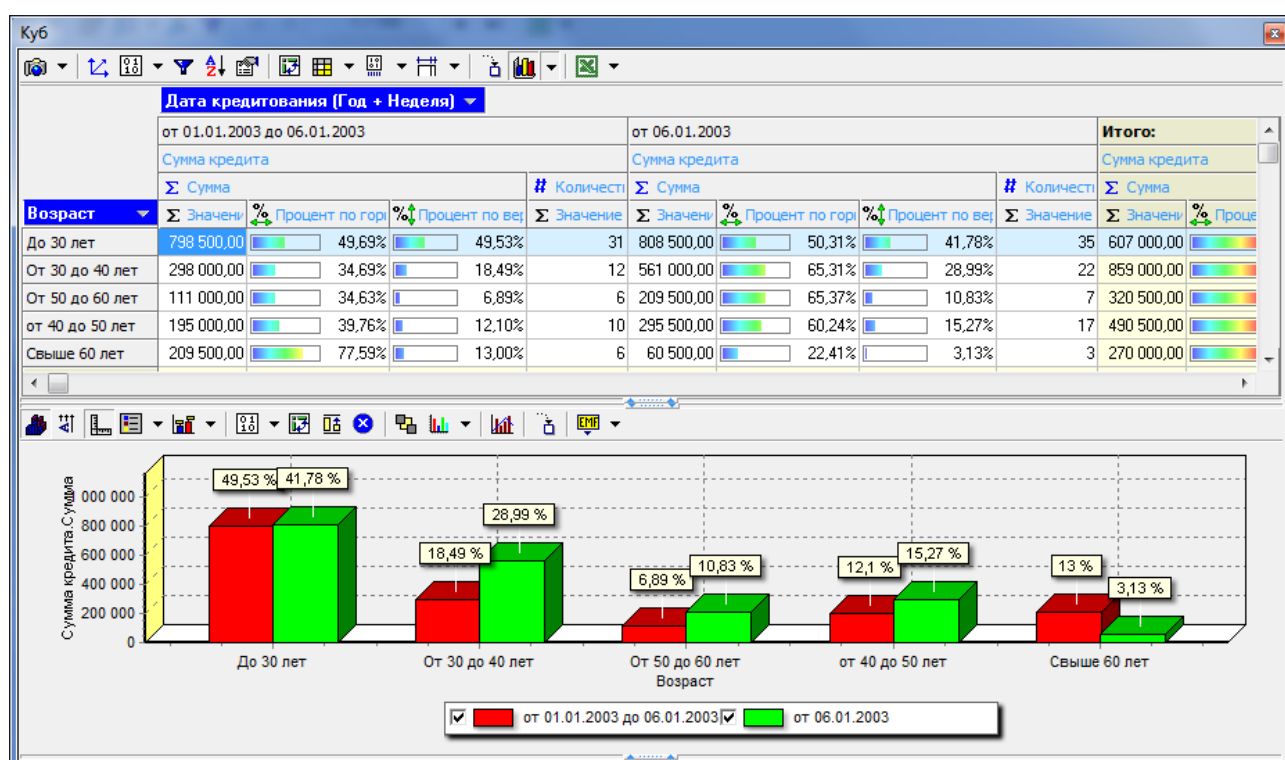


Рис. 33. Информация о кредитовании по возрастным группам и неделям

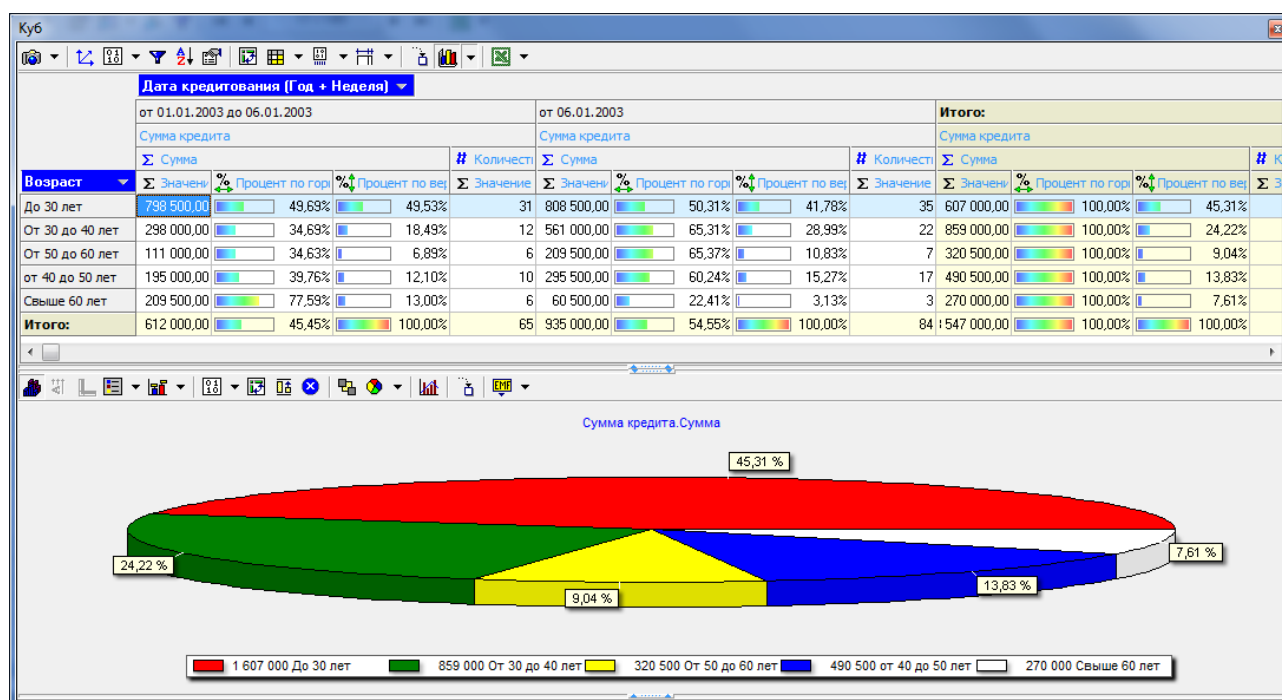


Рис. 34. Итоговая информация о кредитовании по возрастным группам

15. Сохранить проект в личной папке.

4. Замена данных в таблице




Обработчик **Замена данных** предназначен для замены значений набора данных по таблице подстановок, которая содержит пары, состоящие из исходного значения и результирующего значения.

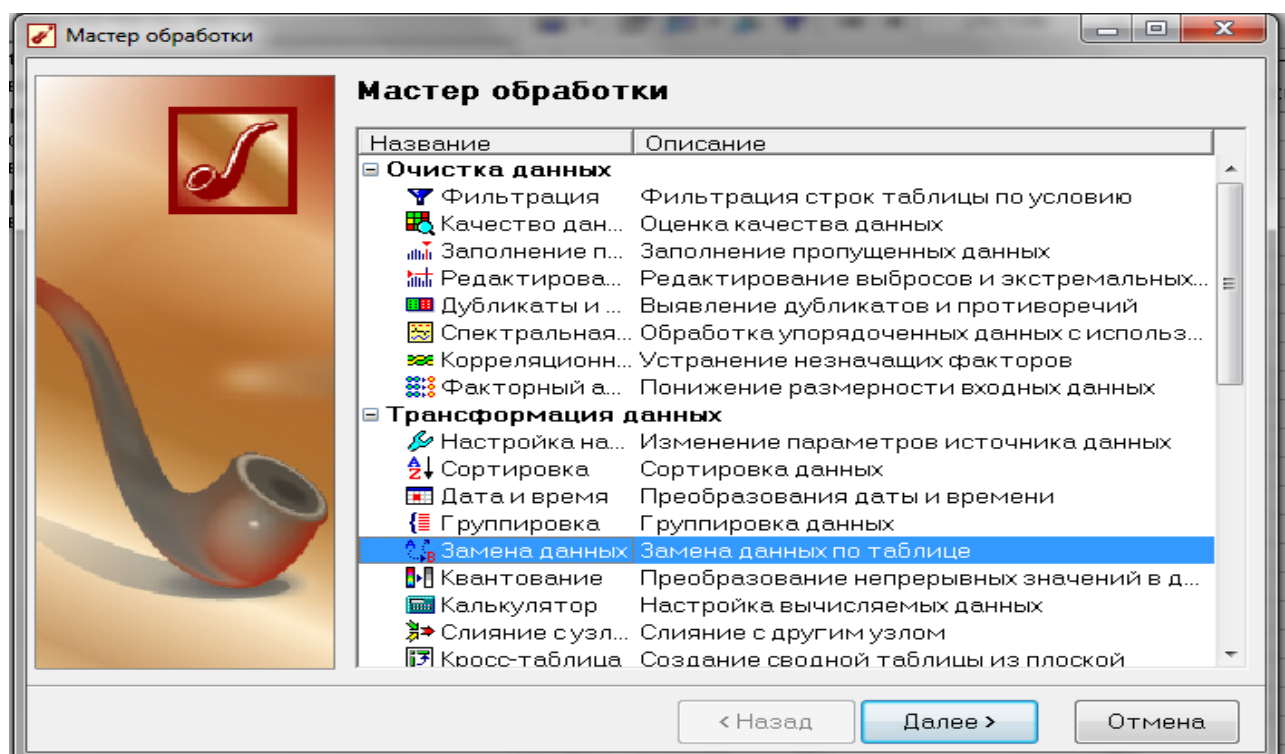
Для каждого значения исходного набора данных ищется соответствие среди исходных значений таблицы подстановки. Если соответствие найдено, то значение меняется на соответствующее выходное значение из таблицы подстановки. Если значение не найдено в таблице, оно может быть либо заменено значением, указанным для замены «по умолчанию», либо оставлено без изменений (если такое значение не указано).

В результате замены для каждого поля, которое в нем участвует, создается новое поле с префиксом **_REPLACE** как к имени, так и к метке поля. Например, для поля *Образование* после узла **Замена данных** появится новое поле *Образование_REPLACE*.

Исходные данные: текстовый файл **Credit.txt**.

Требуется: заменить значения в таблице для поля *Образование*.

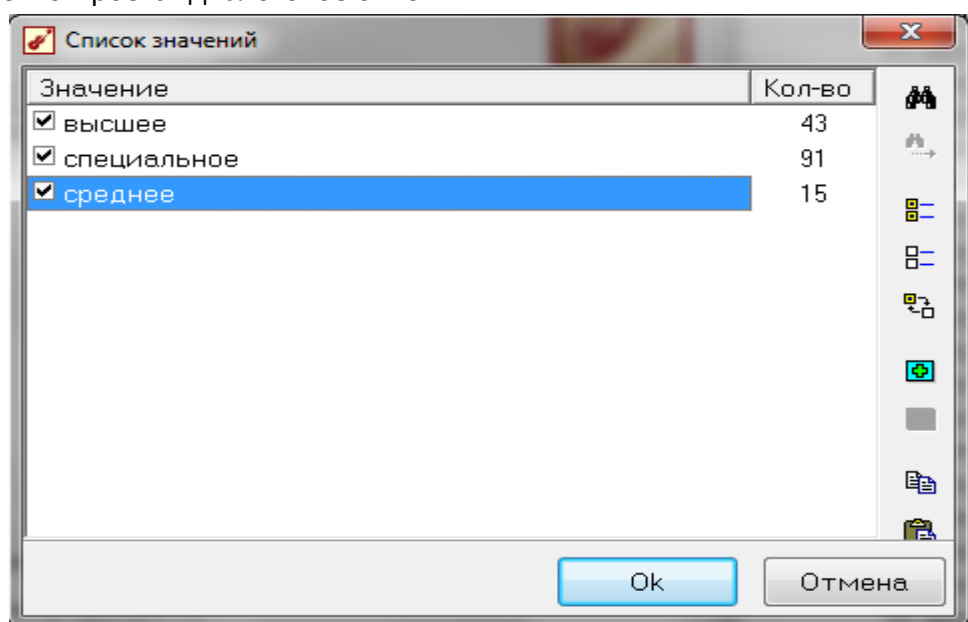
1. Запустить Мастер обработки. В группе Трансформация данных выбрать метод обработки Замена данных (рис.). Перейти к следующему шагу.
2. В окне настройки параметров замены для поля Образование ввести таблицу подстановок. Добавление новой строки в таблицу подстановок производится нажатием кнопки , удаление существующей – . Также ввести все значения списка можно используя кнопку .



В таблице подстановок должны быть заполнены два поля:

- ✓ **Значение** – заменяемое значение поля исходной таблицы. Если поле дискретное, то для ввода значения можно воспользоваться кнопкой выбора, где флажками отметить нужные значения.

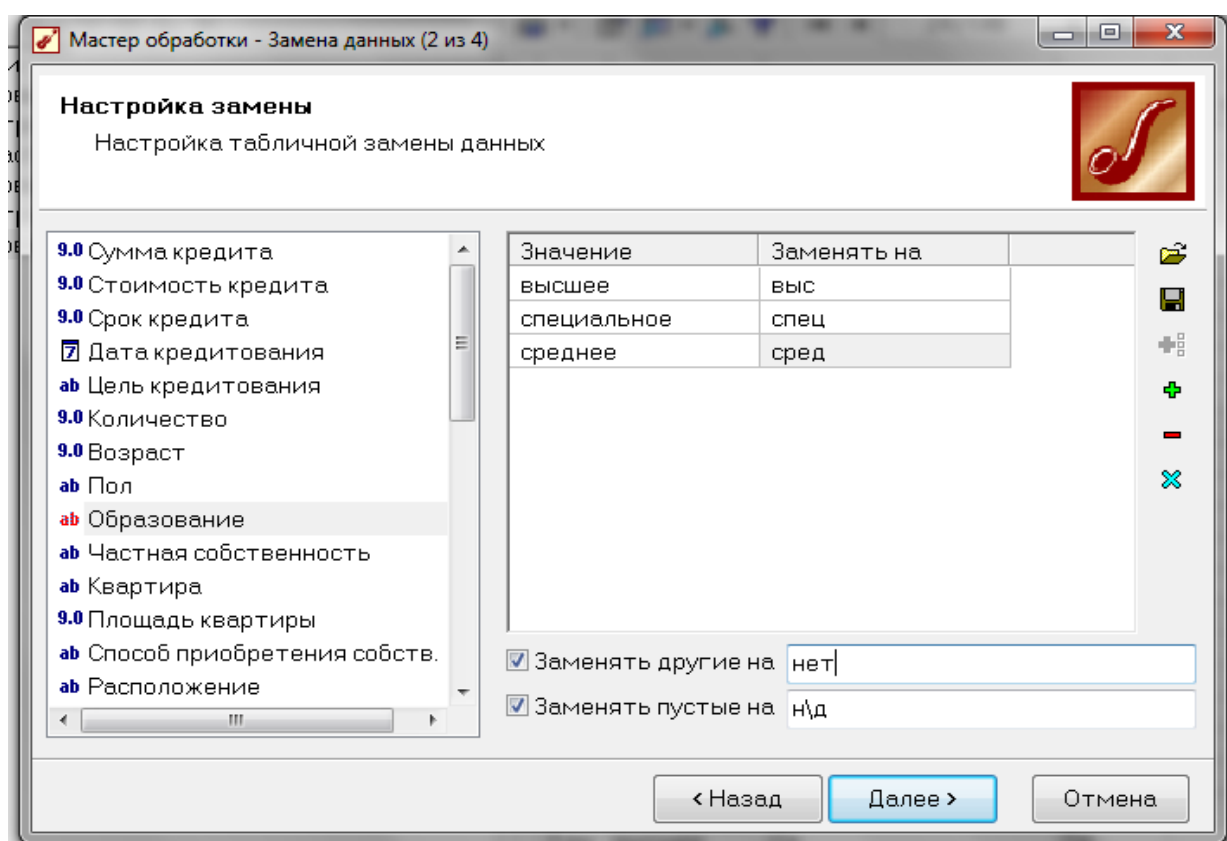
При этом откроется диалоговое окно:



- ✓ **Заменять на** – значение для замены того, что указано в поле *Значение*.

Внизу таблицы подстановок расположены еще два параметра, которые при необходимости можно задать:

- ✓ **Заменять другие на** – на какое значение следует заменить значения, не указанные в таблице замены. Для этого установите флажок и в поле напротив введите значение для замены.
- ✓ **Заменять пустые на** – на какое значение заменять пустые значения поля.



3. В результате в таблице в конце появится новое поле Образование_REPLACE.