ЛАБОРАТОРНАЯ РАБОТА 7 УЗЕЛЫ КВАНТОВАНИЕ, КРОСС – ТАБЛИЦА, ПРЕОБРАЗОВАНИЕ ДАННЫХ К СКОЛЬЗЯЩЕМУ ОКНУ

1. Квантование

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) для анализа или фильтрации, исходя именно из этих интервалов. Для этого 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 записей). В качестве значений результирующего набора данных могут выступать номер интервала, нижняя или верхняя граница интервала, середина интервала либо метка интервала (значения определяемые аналитиком).

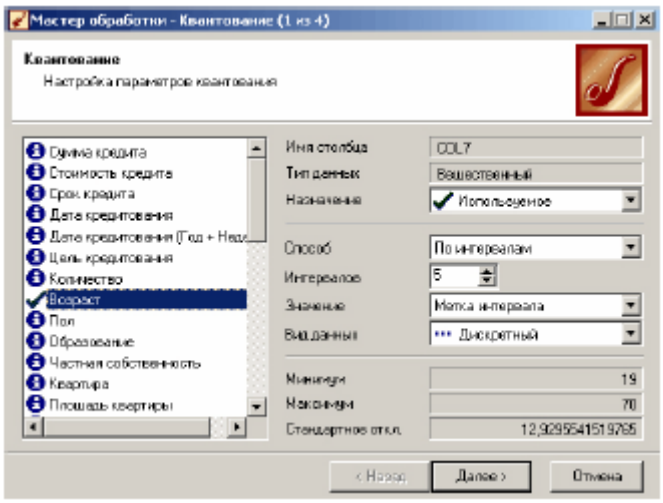
Примером использования данного инструмента может служить разбиение данных о возрасте кредиторов на 5 интервалов (до 30 лет, от 30 до 40, от 40 до 50, от 50 до 60, старше 60 лет). Исходные данные распределятся по пяти интервалам именно так, поскольку согласно статистике минимальное значение возраста кредитора 19, а максимальное - 69 лет. Это необходимо аналитику для оценки кредиторской активности разных возрастных групп с целью принятия решения о стимулировании кредиторов в группах с низкой активностью (например, уменьшение стоимости кредита для этих групп) и, быть может, увеличение прибыли в возрастных группах кредиторов с высоким риском (путем предложения дополнительных **платных** услуг). Причем аналитик желает видеть данные в разрезе по неделям.

Исходные данные

Воспользуемся данными, полученными при разбиении даты файла "Credit.txt".

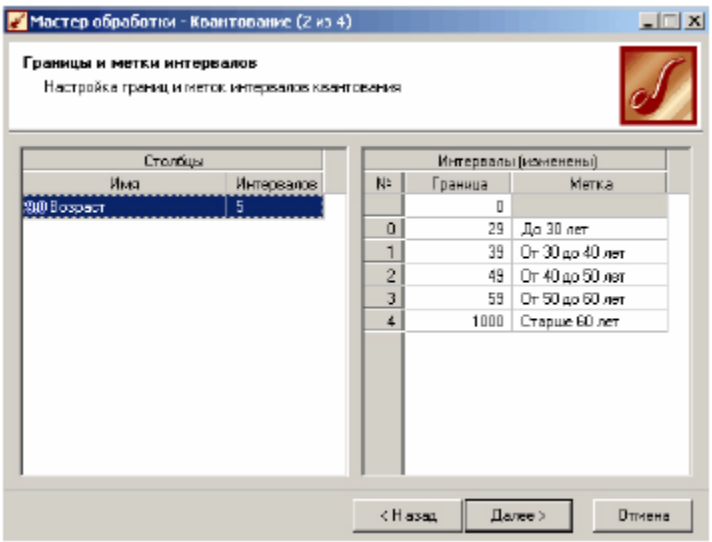
Обработка данных

Воспользуемся Мастером квантования.



В нем выберем назначение поля "Возраст" используемым, укажем способ разбиения "По интервалам", зададим количество интервалов равное 5, в качестве значения выберем "Метка интервала".

На следующем шаге Мастера определим сами метки соответственно возраста кредиторов: "до 30 лет", "от 30 до 40 лет" и т.д.

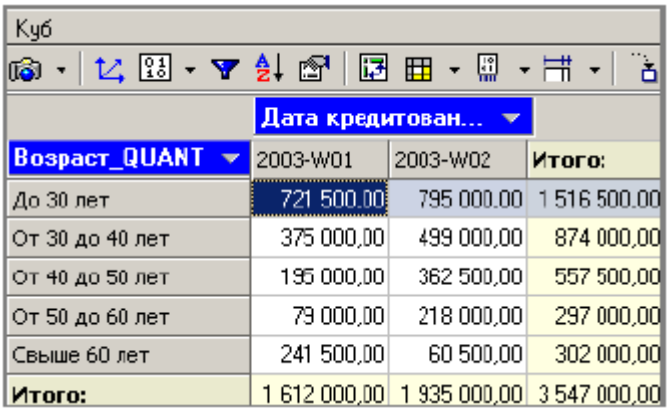


Результат

После обработки выберем в качестве способа отображения "Куб". В Мастере укажем "Сумма кредита" в качестве факта, "Возраст" и поле "Дата кредитования (Год + Неделя)" в качестве измерения, остальные поля укажем неиспользуемыми.

Далее перенесем "Возраст" из доступных измерений в "Измерения в строках", а "Дата кредитования (Год + Неделя)" в "Измерения в столбцах".

На кросс-диаграмме теперь видна информация о том, какие суммы кредитов берут кредиторы определенных возрастных групп по неделям.



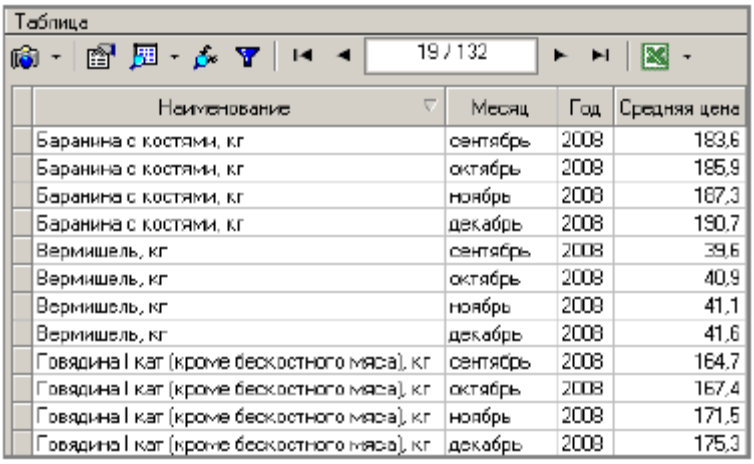
2. Кросс - таблица

Данный обработчик предназначен для преобразования исходной структуры таблицы данных в удобную для работы форму. С его помощью задаются новые поля таблицы из уже существующих, на основе преобразования значений выбранного поля в новые поля с помощью встроенного обработчика фильтрации. Например: поле "месяц" содержащее в себе значения: "январь", "февраль", "март" преобразуется в три соответствующих поля. Значениями которого будут являться агрегированное поле фактов заданное аналитиком. Данный обработчик можно заменить обработчиками: "Фильтр" - с помощью которого выбираются значения на основе которых будет строится первое поле таблицы, далее применяется "Калькулятор" - который формирует измерения нового поля и присваивает ему имя; данный алгоритм повторяется для всех предусмотренных полей; после чего все созданные поля собирают с помощью "Группировки" .

На основе кросс - таблицы удобно вычислять экономические показатели рассчитываемые на основе прошедших месяцев. "Кросс - таблица" является одним из инструментов **Deductor Studio.**

Исходные данные

Продемонстрируем применение "Кросс-таблицы", используя данные о стоимости продуктов входящих в потребительскую корзину за четыре месяца. Исходные данные находятся в файле "basket_of_goods.txt". Необходимо оценить индексы роста цен на продукты питания.



Таблица

19 / 132

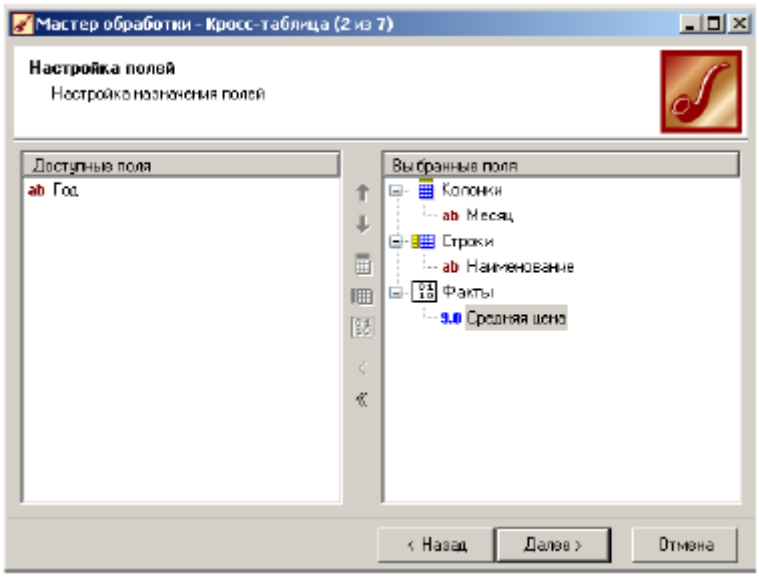
Наименование	Месяц	Год	Средняя цена
Баранина с костями, кг	сентябрь	2008	183,6
Баранина с костями, кг	октябрь	2008	185,9
Баранина с костями, кг	ноябрь	2008	187,3
Баранина с костями, кг	декабрь	2008	190,7
Вермишель, кг	сентябрь	2008	39,6
Вермишель, кг	октябрь	2008	40,9
Вермишель, кг	ноябрь	2008	41,1
Вермишель, кг	декабрь	2008	41,6
Говядина I кат (кроме бескостного мяса), кг	сентябрь	2008	164,7
Говядина I кат (кроме бескостного мяса), кг	октябрь	2008	167,4
Говядина I кат (кроме бескостного мяса), кг	ноябрь	2008	171,5
Говядина I кат (кроме бескостного мяса), кг	декабрь	2008	175,3

Вид исходной таблицы мало пригоден для вычислений индексов. Данную таблицу необходимо отредактировать, что бы в ней появились дополнительные поля. В которых содержалась бы информация о цене рассматриваемых продуктов питания за каждый месяц в отдельности. Применим обработчик "Кросс-таблица".

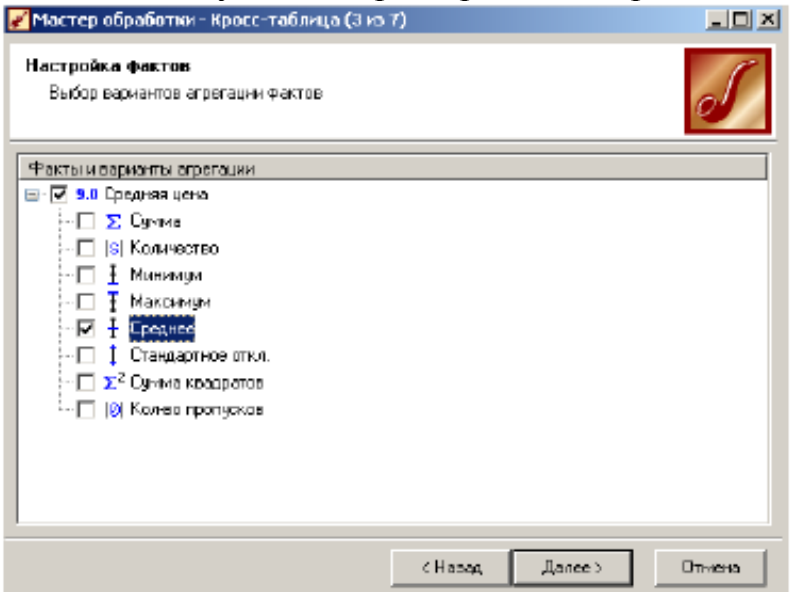
Преобразование исходной таблицы данных

Вызовем "Мастер обработки" и в появившемся окне выберем обработчик "Кросс - таблица".

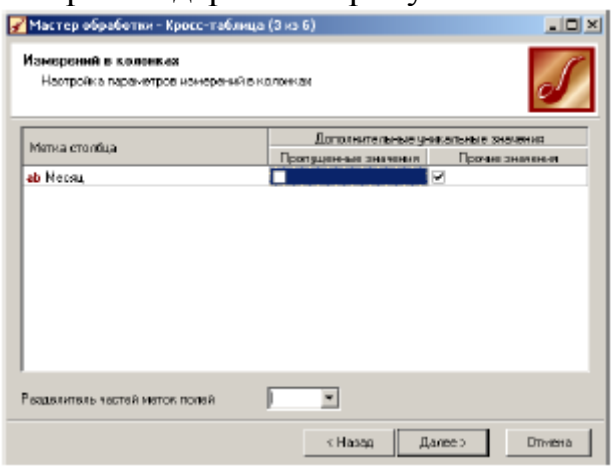
Следующим шагом будет настройка используемых полей для формирования таблицы. Используемые поля для построения должны находится либо в колонках либо в строках. В колонки помещают поля на основе значений которых будут создаваться новые, их значениями будут выбранные факты. В строки помещаются поля, которые не нуждаются в изменении. Настроим данное окно: переместим "Месяц" в колонки, а "Наименование" в строки, при этом необходимо обязательно указать факты в данном случае - "Средняя цена". Новая таблица будет содержать поля: "Наименование" - название продуктов входящих в потребительскую корзину; "Сентябрь" - средняя цена, данных продуктов за сентябрь месяц, ... "Декабрь" - средняя цена, продуктов за декабрь месяц.



Следующим шагом необходимо настроить параметры агрегации выбранных фактов. В нашем случае выберем среднюю агрегацию.



После нажатия кнопки "Далее" открывается следующее окно "Мастера обработки", в котором выбирается настройка параметров измерений в колонках. В нем резервируются дополнительные поля для возможного внесения изменений в значения исходного поля таблицы, а так же для измерений, в названии которых содержатся пропуски.

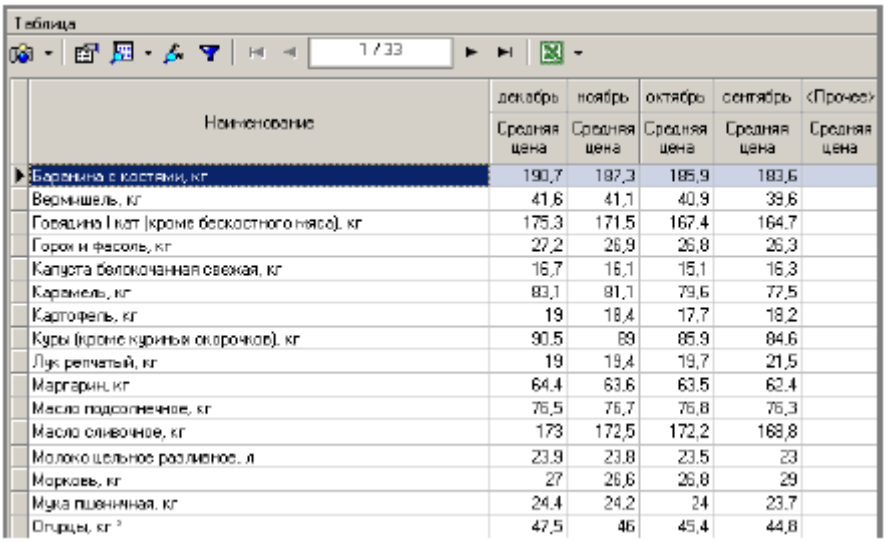


Так как у нас нет данных о цене товара, с неопределенным месяцем, то галочку рядом с "Пропущенными значениями" ставить не будем. "Прочие значения" отметим галочкой, так как в дальнейшем мы рассчитываем пополнить исходную таблицу еще одним месяцем, данные которого запишутся в данный столбец.

Все настройки заданы, запустим процесс на выполнение.

Результат

Из множества предлагаемых визуализаторов выберем " Таблицу"



Таблица

1 / 33

Наименование	декабрь	ноябрь	октябрь	сентябрь	<Прочее>
	Средняя цена	Средняя цена	Средняя цена	Средняя цена	Средняя цена
Баранина с костями, кг	190,7	187,3	185,9	183,6	
Вермишель, кг	41,6	41,1	40,9	39,6	
Говядина I кат (кроме бескостного мяса), кг	175,3	171,5	167,4	164,7	
Горох и фасоль, кг	27,2	26,9	26,8	26,3	
Капуста белокочанная свежая, кг	16,7	16,1	15,1	16,3	
Карамель, кг	83,1	81,1	79,6	77,5	
Картофель, кг	19	18,4	17,7	18,2	
Куры (кроме куриных окорочков), кг	90,5	89	85,9	84,6	
Лук репчатый, кг	19	19,4	19,7	21,5	
Маргарин, кг	64,4	63,6	63,5	62,4	
Масло подсолнечное, кг	76,5	76,7	76,8	76,3	
Масло сливочное, кг	173	172,5	172,2	168,8	
Молоко цельное разливное, л	23,9	23,8	23,5	23	
Морковь, кг	27	26,6	26,8	29	
Мука пшеничная, кг	24,4	24,2	24	23,7	
Огурцы, кг	47,5	46	45,4	44,8	

Таким образом, после обработки получили новую таблицу данных, на основе которой удобно производить необходимые вычисления индексов.

Данную таблицу можно получить с помощью группы обработчиков: "Фильтр", "Калькулятор" и "Группировка", но они делают сценарий очень громоздким и неудобным к исправлению. Использование "Кросс-диаграммы" существенно сокращает время построения сценария и обработки.

3. Преобразование данных к скользящему окну

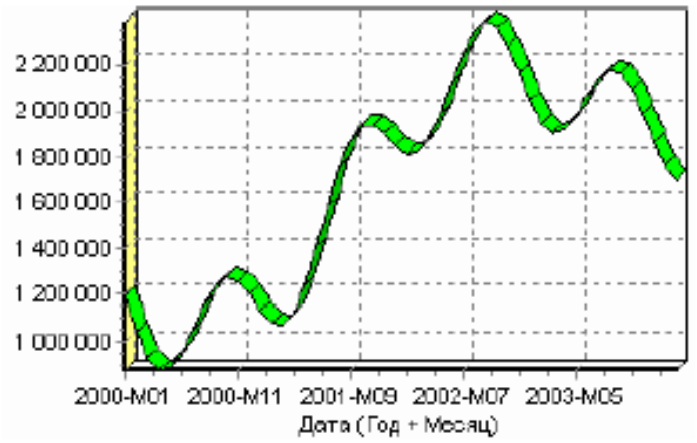
Когда требуется прогнозировать временной ряд, тем более, если налицо его периодичность (сезонность), то лучшего результата можно добиться, учитывая значения факторов не только в данный момент времени, но и, например, за аналогичный период прошлого года. Такую возможность можно получить после трансформации данных к скользящему окну. То есть, например, при сезонности продаж с периодом 12 месяцев, для прогнозирования количества продаж на месяц вперед можно в качестве входного фактора указать не только значение количества продаж за предыдущий месяц, но и за 12 месяцев назад.

Обработка создает новые столбцы путем сдвига данных исходного столбца вниз и вверх (глубина погружения и горизонт прогноза).

Исходные данные

У аналитика имеются данные о месячном количестве проданного товара за несколько лет. Ему необходимо, основываясь на этих данных, сказать, какое количество товара будет продано через неделю и через две.

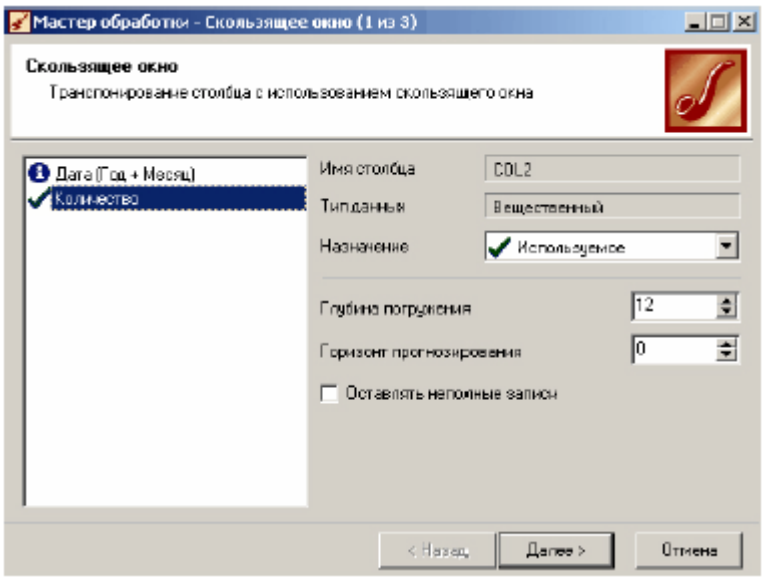
Исходные данные по продажам находятся в файле "Trade.txt". Выполним импорт данных из файла, не забыв указать в Мастере, чтобы в качестве разделителя дробной и целой части была точка, а не запятая. Выполнив удаление аномалий и сглаживание, получаем:



Приведение данных к скользящему окну

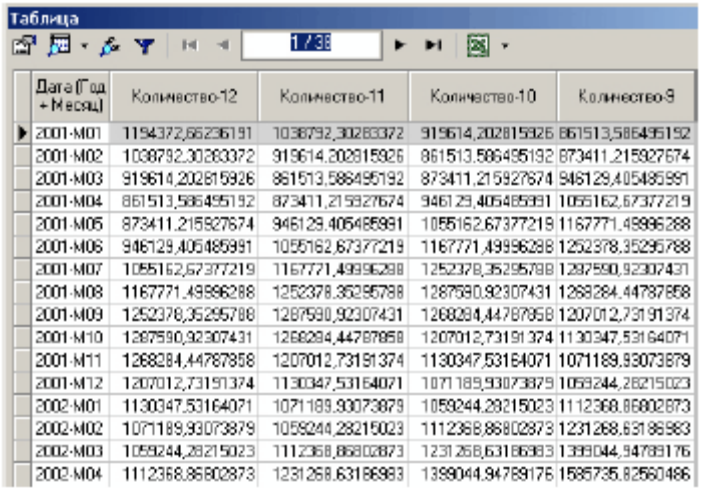
Запустим Мастер обработки, выберем в качестве обработчика скользящее окно и перейдем на следующий шаг.

Можно использовать обработчик "Автокорреляция" и убедиться в наличии годовой сезонности. В связи с этим строить прогноз на месяц вперед можно, основываясь на данных за 1, 2, 11 и 12 месяцев назад. Поэтому необходимо, назначив поле "Количество" используемым, выбрать глубину погружения 12. Тогда данные трансформируются к скользящему окну так, что аналитику будут доступны все требуемые факторы для построения прогноза.



Результат

Просмотреть полученные данные можно в виде таблицы.



Таблица

1 / 38

Дата (Год + Месяц)	Количество-12	Количество-11	Количество-10	Количество-9
2001-M01	1194372,66236191	1038792,30283372	919614,202815926	861513,5864955152
2001-M02	1038792,30283372	919614,202815926	861513,5864955192	873411,215927674
2001-M03	919614,202815926	861513,5864955192	873411,215927674	946129,405485991
2001-M04	861513,5864955192	873411,215927674	946129,405485991	1055162,67377219
2001-M05	873411,215927674	946129,405485991	1055162,67377219	1167771,49996288
2001-M06	946129,405485991	1055162,67377219	1167771,49996288	1252378,35295788
2001-M07	1055162,67377219	1167771,49996288	1252378,35295788	1287590,92307431
2001-M08	1167771,49996288	1252378,35295788	1287590,92307431	1268284,44787858
2001-M09	1252378,35295788	1287590,92307431	1268284,44787858	1207012,73191374
2001-M10	1287590,92307431	1268284,44787858	1207012,73191374	1130347,53164071
2001-M11	1268284,44787858	1207012,73191374	1130347,53164071	1071189,93073879
2001-M12	1207012,73191374	1130347,53164071	1071189,93073879	1059244,28215023
2002-M01	1130347,53164071	1071189,93073879	1059244,28215023	1112368,86802873
2002-M02	1071189,93073879	1059244,28215023	1112368,86802873	1231268,63186983
2002-M03	1059244,28215023	1112368,86802873	1231268,63186983	1399044,94789176
2002-M04	1112368,86802873	1231268,63186983	1399044,94789176	1585735,82560486

Как видно, теперь в качестве входных факторов можно использовать "Количество - 12", "Количество - 11" - данные по количеству 12 и 11 месяцев назад (относительно прогнозируемого месяца) и остальные необходимые факторы. В качестве результата прогноза будет указан столбец "Количество".

Лабораторная работа:

1. Выполните **Квантование** данных из файла "Credit.txt"
2. Выполните преобразование данных файла "basket_of_goods.txt"с помощью обработчика **Кросс таблица**
3. Выполните **Преобразование данных к скользящему окну** из файла "Trade.txt".

Вопросы для проверки:

1. Назначения и ход выполнения **Квантования** данных
2. Назначение обработчика **Кросс таблица**, порядок выполнения
3. Назначение обработчика **Преобразование данных к скользящему окну**, ход выполнения