



МЕТОДЫ ОЧИСТКИ ДАННЫХ: ЧАСТОТНЫЙ АНАЛИЗ

gender.lgd

#	Имя	Пол
1	Ольга	Ж
2	Константин	М
3	Юлия	Ж
4	Любовь	Ж
5	Анна	Ж
6	Елена	Ж
7	Иван	М
8	Лариса	М
9	Татьяна	Ж
10	Екатерина	Ж
11	Виктор	М
12	Николай	М
13	Максим	М
14	Сергей	М
15	Наталья	Ж
16	Денис	М
17	Ольга	Ж
128	Елена	Ж

Таблица Форма

Для проведения частотного анализа достаточно, узла кросс-таблицы. Продemonстрируем это на примере, который рассматривали в лекции: импортируем в новом модуле нашего пакета набор данных **gender.lgd**, где содержатся поля **Имя** и **Пол**.

gender.lgd



Кросс-таблица



Добавим узел **Кросс-таблица** в сценарий и настроим его.

Фильтрация

Доступные поля

В колонки поместим поле **Пол**, в строки – **Имя**, а в факты добавим **Количество** с видом агрегации **Количество**.

Разделитель частей меток полей

Скользящие уникальные значения

Ограничение количества значений

Уникальные значения как имена полей



Выбранные поля

Колонки

ab Пол

Строки

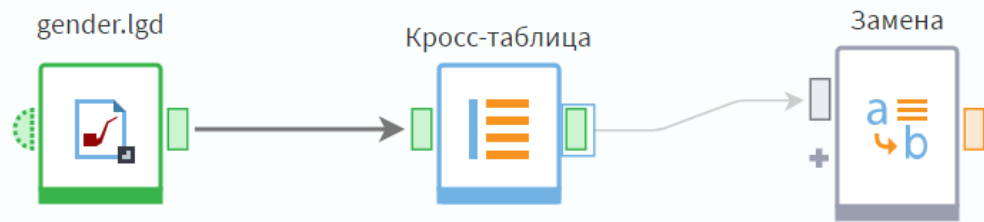
ab Имя

Σ Факты

12 Количество (Количество)

Получили частотную таблицу распределения. Всего в наборе оказалось 16 имен. Составим справочник соответствий **Имя-Пол**.

#	ab Имя	12 Ж	12 М
1	Анна	12	1
2	Виктор	1	8
3	Денис		5
4	Екатерина	4	2
5	Елена	11	2
6	Иван	1	11
7	Константин	1	7
8	Лариса	4	3
9	Любовь	8	
10	Максим		7
11	Наталья	4	2
12	Николай		5
13	Ольга	5	
14	Сергей	3	7
15	Татьяна	5	
16	Юлия	9	



Добавим узле **Замена**.

Поля	Способ замены
ab Имя	Не заменять
12 Ж	Ввод вручную
12 М	Ввод вручную

Импорт Экспорт Сортировать 12 Изменить тип замены Редактировать Получить значения

#	12 Значение	12 Замена	
	Точное совпадение		+
1		0	+

Для обоих полей выберем способ замены **Ввод вручную** и добавим замену по точному совпадению на 0.



Далее воспользуемся узлом **Калькулятор**.

Выражения *f(x)* |

Имя	Метка
ab Gender	Пол

Поля/Переменные

Имя	Метка
Поля	
ab FirstName	Имя
12 C_1	Ж Замена
0/1 C_1_Replaced	Ж Заменен
12 C_2	М Замена
0/1 C_2_Replaced	М Заменен

Предпросмотр... | AND OR NOT XOR | = <> < > <= >= | 9.0 "" 31

```
If(Max(C_1,C_2)=C_1,"Ж","М")
```

В его настройках зададим выражение, как показано на слайде. Оно означает: если число в поле **Ж** больше числа в поле **М**, значение пола – **Ж**, иначе – **М**.

Список

- 9.0 AbsErr (Аргумент1, Аргумент2)
- 31 AddDay (Дата, Количество)
- 31 AddMonth (Дата, Количество)
- 31 AddQuarter (Дата, Количество)
- 31 AddWeek (Дата, Количество)
- 31 AddYear (Дата, Количество)
- 9.0 AMGD (Стоимость, Остаточная_стоимость, Время...

Получили список корректных значений. При необходимости их можно присоединить к исходному списку, а также сохранить в качестве справочника.

#	ab Пол	ab Имя	12 Ж Замена	0/1 Ж Заменен	12 М Заме...	0/1 М Замен...
1	Ж	Анна	12	false	1	false
2	М	Виктор	1	false	8	false
3	М	Денис	0	true	5	false
4	Ж	Екатерина	4	false	2	false
5	Ж	Елена	11	false	2	false
6	М	Иван	1	false	11	false
7	М	Константин	1	false	7	false
8	Ж	Лариса	4	false	3	false
9	Ж	Любовь	8	false	0	true
10	М	Максим	0	true	7	false
11	Ж	Наталья	4	false	2	false
12	М	Николай	0	true	5	false
13	Ж	Ольга	5	false	0	true
14	М	Сергей	3	false	7	false
15	Ж	Татьяна	5	false	0	true
16	Ж	Юлия	9	false	0	true