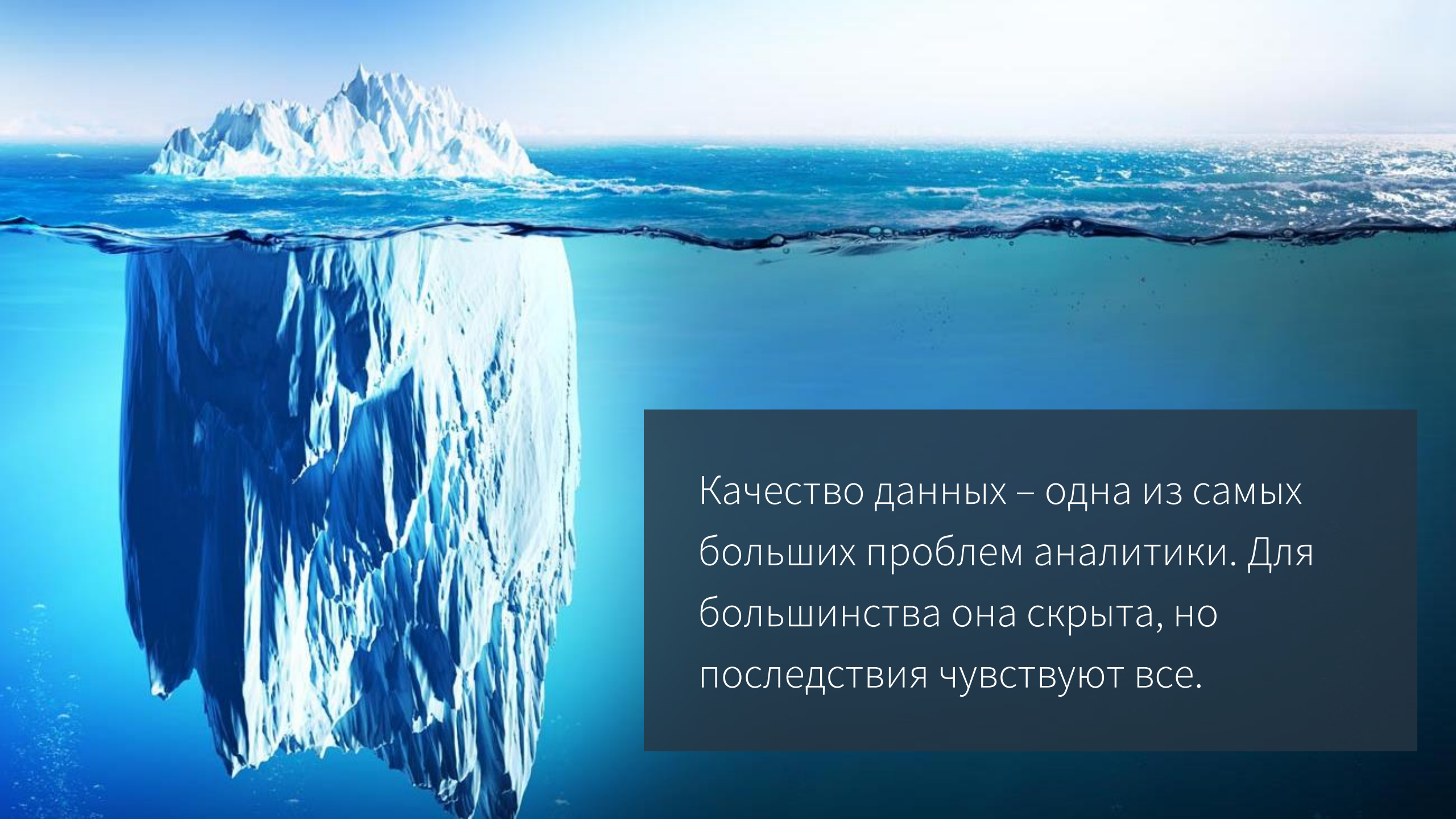
Визуализация качества данных в Loginom

Арустамов Алексей



Качество данных – одна из самых больших проблем аналитики. Для большинства она скрыта, но последствия чувствуют все.

Критерии качества

1. Согласованность — взаимная непротиворечивость
2. Своевременность — доступность в нужный момент
3. Целостность — корректность ссылок, соответствие правилам
4. Точность — нужный уровень детализации
5. Полнота — достаточность объема, глубины и широты
6. Правильность — отсутствие систематических ошибок
7. …

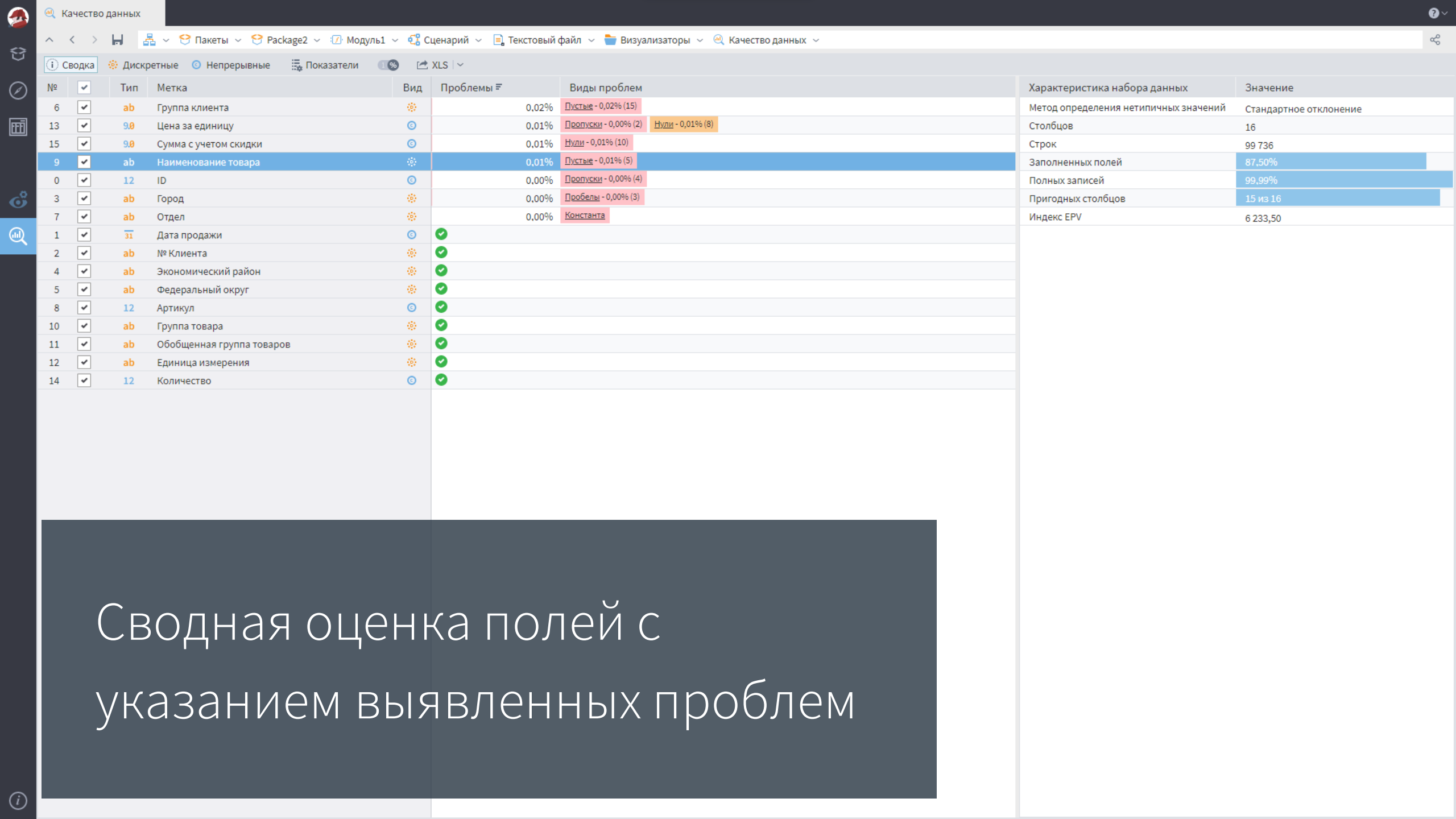
Способы повышения качества

1. **Профилирование** – исследование данных для выяснения их статистических характеристик, таких как распределения величин, наличие выбросов, параметры выборки
2. **Верификация** – проверка данных по критериям достоверности источника, точности, согласованности и соответствия формату
3. **Стандартизация** – приведение к определённому формату и представлению, который обеспечивает корректный анализ

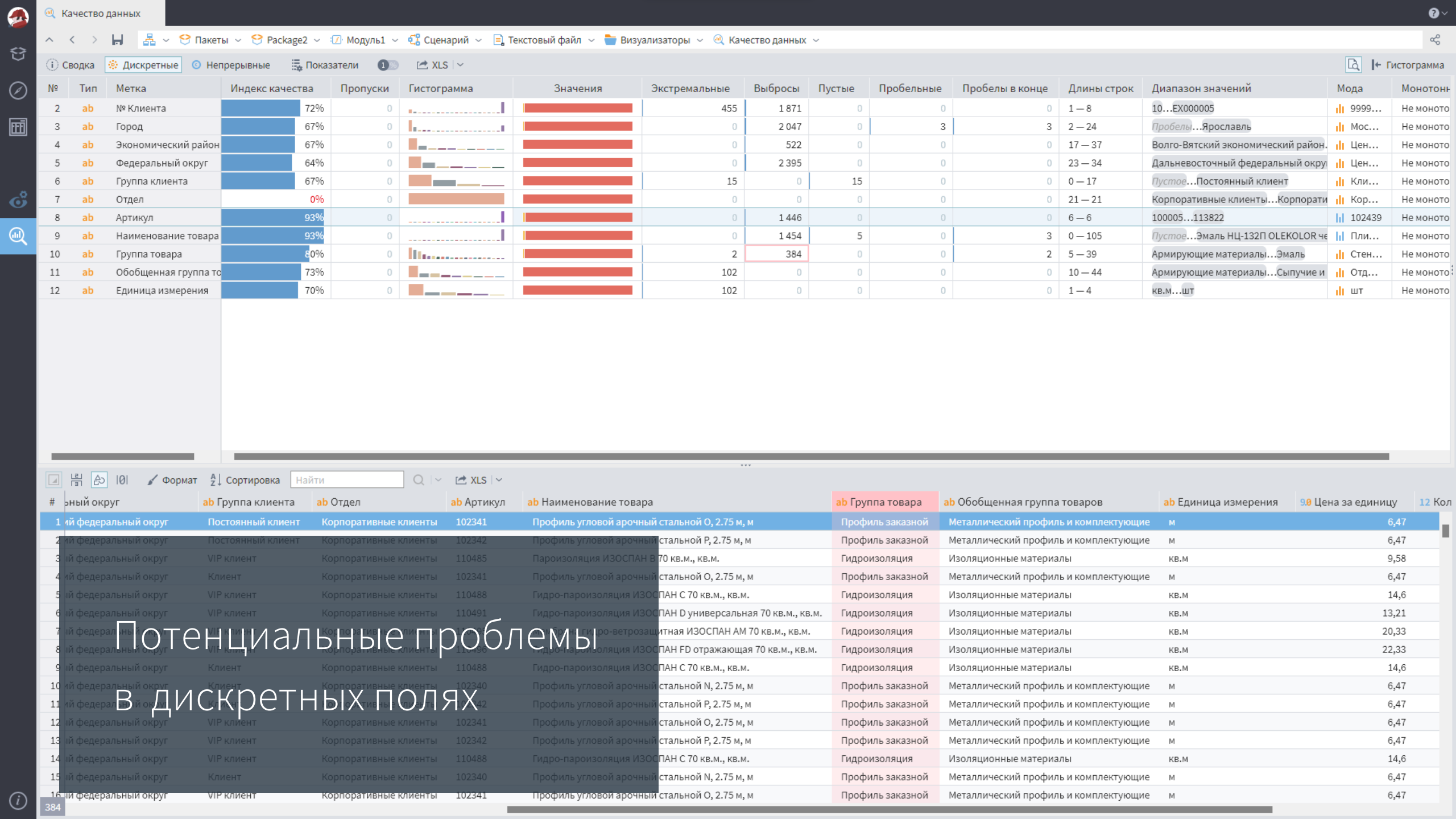
Визуализатор «Качество данных»

Расчет характеристик выборок для выявления типичных индикаторов проблем в данных:

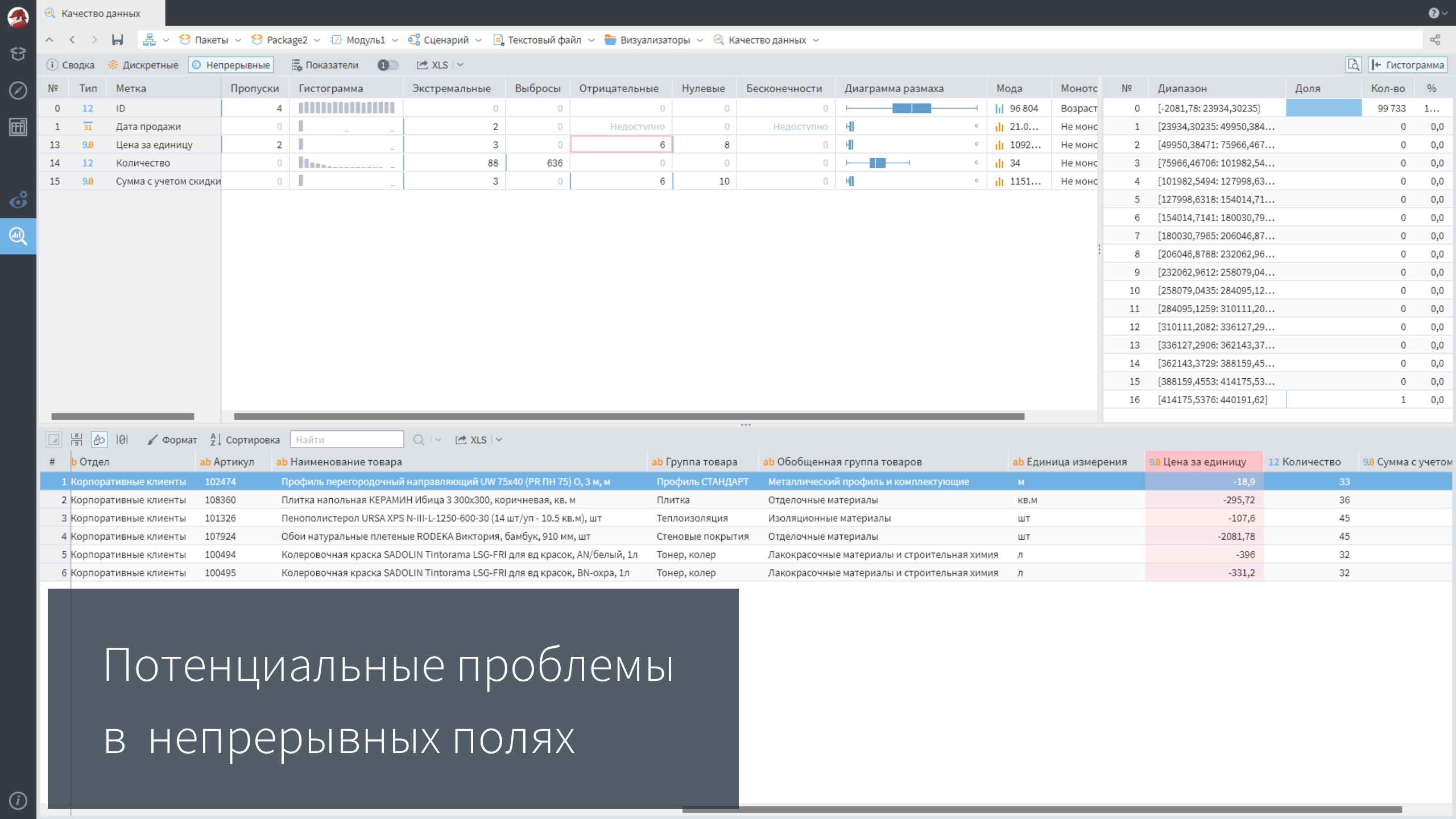
1. Пропуски и пробельные значения
2. Выбросы и экстремальные значения
3. Наличие моды и монотонность
4. Распределения значений
5. ….

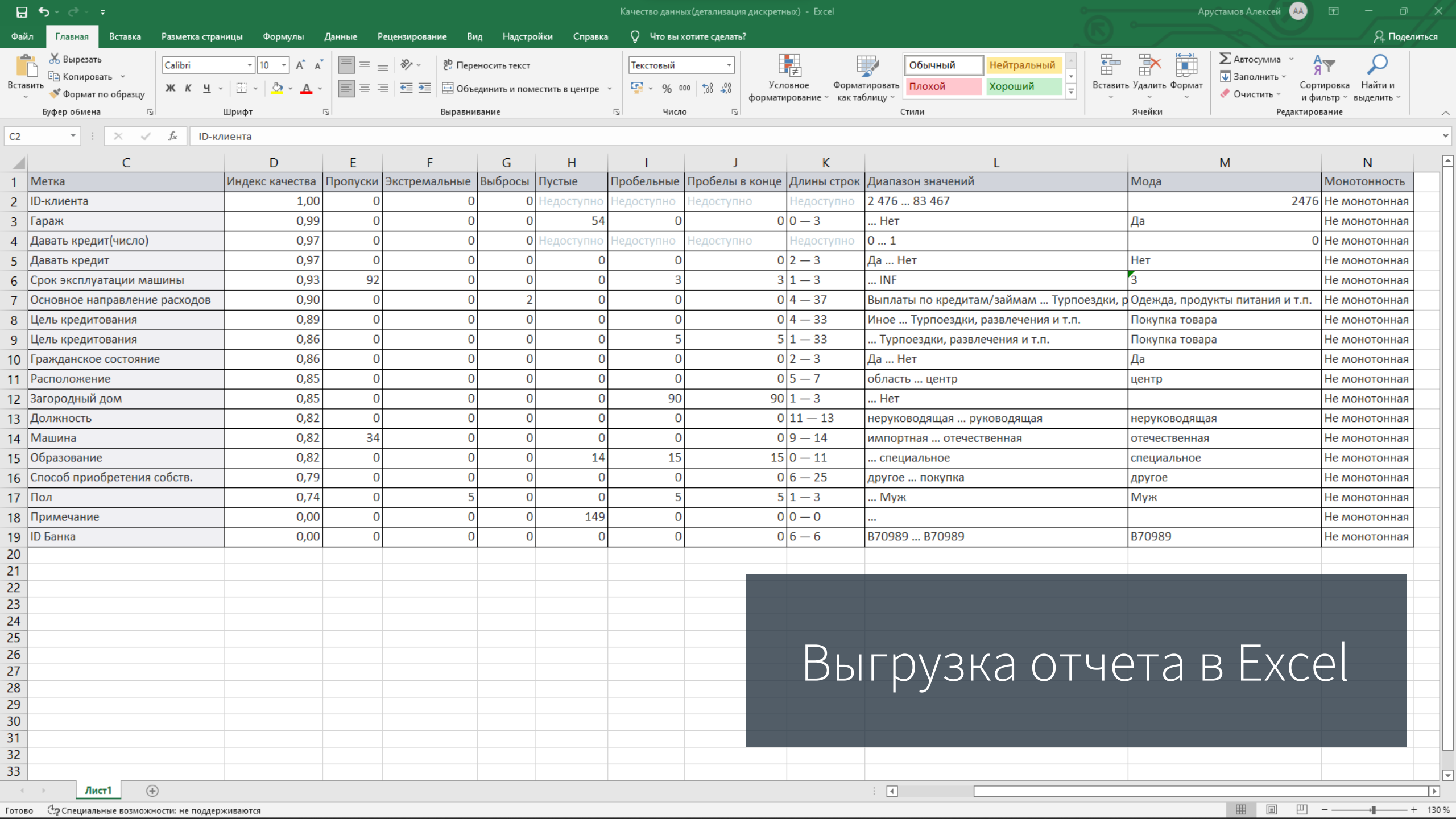


Сводная оценка полей с указанием выявленных проблем



Потенциальные проблемы в дискретных полях


Качество данных
Пакеты
Package2
Модуль1
Сценарий
Текстовый файл
Визуализаторы
Качество данных
Сводка
Дискретные
Непрерывные
Показатели
XLS
Гистограмма
№ Тип Метка Пропуски Гистограмма Экстремальные Выбросы Отрицательные Нулевые Бесконечности Диаграмма размаха Мода Монотс
0 12 ID 4 0 0 0 0 0 96 804 Возраст
1 31 Дата продажи 0 2 0 Недоступно 0 Недоступно 21.0… Не монс
13 9.0 Цена за единицу 2 3 0 6 8 0 1092… Не монс
14 12 Количество 0 88 636 0 0 0 34 Не монс
15 9.0 Сумма с учетом скидки 0 3 0 6 10 0 1151… Не монс
№ Диапазон Доля Кол-во %
0 [-2081,78: 23934,30235) 99 733 1…
1 [23934,30235: 49950,384… 0 0,0
2 [49950,38471: 75966,467… 0 0,0
3 [75966,46706: 101982,54… 0 0,0
4 [101982,5494: 127998,63… 0 0,0
5 [127998,6318: 154014,71… 0 0,0
6 [154014,7141: 180030,79… 0 0,0
7 [180030,7965: 206046,87… 0 0,0
8 [206046,8788: 232062,96… 0 0,0
9 [232062,9612: 258079,04… 0 0,0
10 [258079,0435: 284095,12… 0 0,0
11 [284095,1259: 310111,20… 0 0,0
12 [310111,2082: 336127,29… 0 0,0
13 [336127,2906: 362143,37… 0 0,0
14 [362143,3729: 388159,45… 0 0,0
15 [388159,4553: 414175,53… 0 0,0
16 [414175,5376: 440191,62] 1 0,0
Формат
Сортировка
Найти
XLS
Отдел Артикул Наименование товара Группа товара Обобщенная группа товаров Единица измерения Цена за единицу Количество Сумма с учетом
1 Корпоративные клиенты 102474 Профиль перегородочный направляющий UW 75x40 (PR ПН 75) О, 3 м, м Профиль СТАНДАРТ Металлический профиль и комплектующие м -18,9 33
2 Корпоративные клиенты 108360 Плитка напольная КЕРАМИН Ибица 3 300x300, коричневая, кв. м Плитка Отделочные материалы кв.м -295,72 36
3 Корпоративные клиенты 101326 Пенополистерол URSA XPS N-III-L-1250-600-30 (14 шт/уп - 10.5 кв.м), шт Теплоизоляция Изоляционные материалы шт -107,6 45
4 Корпоративные клиенты 107924 Обои натуральные плетеные RODEKA Виктория, бамбук, 910 мм, шт Стеновые покрытия Отделочные материалы шт -2081,78 45
5 Корпоративные клиенты 100494 Колеровочная краска SADOLIN Tintorama LSG-FRI для вд красок, AN/белый, 1л Тонер, колер Лакокрасочные материалы и строительная химия л -396 32
6 Корпоративные клиенты 100495 Колеровочная краска SADOLIN Tintorama LSG-FRI для вд красок, BN-охра, 1л Тонер, колер Лакокрасочные материалы и строительная химия л -331,2 32
Потенциальные проблемы
в непрерывных полях



	C	D	E	F	G	H	I	J	K	L	M	N
1	Метка	Индекс качества	Пропуски	Экстремальные	Выбросы	Пустые	Пробельные	Пробелы в конце	Длины строк	Диапазон значений	Мода	Монотонность
2	ID-клиента	1,00	0	0	0	Недоступно	Недоступно	Недоступно	Недоступно	2 476 ... 83 467	2476	Не монотонная
3	Гараж	0,99	0	0	0	54	0	0	0 — 3	... Нет	Да	Не монотонная
4	Давать кредит(число)	0,97	0	0	0	Недоступно	Недоступно	Недоступно	Недоступно	0 ... 1	0	Не монотонная
5	Давать кредит	0,97	0	0	0	0	0	0	2 — 3	Да ... Нет	Нет	Не монотонная
6	Срок эксплуатации машины	0,93	92	0	0	0	3	3	1 — 3	... INF	3	Не монотонная
7	Основное направление расходов	0,90	0	0	2	0	0	0	4 — 37	Выплаты по кредитам/займам ... Турпоездки, р	Одежда, продукты питания и т.п.	Не монотонная
8	Цель кредитования	0,89	0	0	0	0	0	0	4 — 33	Иное ... Турпоездки, развлечения и т.п.	Покупка товара	Не монотонная
9	Цель кредитования	0,86	0	0	0	0	5	5	1 — 33	... Турпоездки, развлечения и т.п.	Покупка товара	Не монотонная
10	Гражданское состояние	0,86	0	0	0	0	0	0	2 — 3	Да ... Нет	Да	Не монотонная
11	Расположение	0,85	0	0	0	0	0	0	5 — 7	область ... центр	центр	Не монотонная
12	Загородный дом	0,85	0	0	0	0	90	90	1 — 3	... Нет		Не монотонная
13	Должность	0,82	0	0	0	0	0	0	11 — 13	неруководящая ... руководящая	неруководящая	Не монотонная
14	Машина	0,82	34	0	0	0	0	0	9 — 14	импортная ... отечественная	отечественная	Не монотонная
15	Образование	0,82	0	0	0	14	15	15	0 — 11	... специальное	специальное	Не монотонная
16	Способ приобретения собств.	0,79	0	0	0	0	0	0	6 — 25	другое ... покупка	другое	Не монотонная
17	Пол	0,74	0	5	0	0	5	5	1 — 3	... Муж	Муж	Не монотонная
18	Примечание	0,00	0	0	0	149	0	0	0 — 0	...		Не монотонная
19	ID Банка	0,00	0	0	0	0	0	0	6 — 6	B70989 ... B70989	B70989	Не монотонная

Визуализатор «Качество данных» позволяет:

1. Быстро оценить пригодность выборки для анализа
2. Автоматически провести типовые проверки качества данных
3. Локализовать проблемы

loginom.ru