



Прикладные задачи анализа данных

искусство визуализации

Часть 2. Прикладная

Дьяконов А.Г.

**Московский государственный университет
имени М.В. Ломоносова (Москва, Россия)**

Цели визуализации

- **анализ (часть EDA = Exploratory Data Analysis)**
 - иллюстрация слов
 - рассказ (story-telling)

нахождение закономерностей

детекция наличия выбросов/аномалий

проверка данных на логичность, полноту и т.п.

придумывание признаков (деформаций, комбинаций, индикаторов)

понимание смысла задачи, проверка гипотез и предварительный выбор модели

Надо изучить предметную область!

Надо обязательно смотреть на данные!

I – способы исследования отдельных признаков

**устанавливаем природу признаков
проверяем логичность признаков
для каждого признака**

- имя
- область значений
- распределение
- устойчивость
- важность

если что-то нарушается... пользуемся этим

Что просто визуализировать

- статистики признаков (описательные из МС)
- характеристики признаков (важности, AUC и т.п.)

Описательные статистики – среднее

$$x_1 \leq \dots \leq x_m$$

Выборочное среднее

$$\text{mean}(X) = \frac{x_1 + \dots + x_m}{m}$$

Медиана

$$\text{median}(X) = q_{0.5}(X) = \frac{x_{\lfloor n/2 \rfloor} + x_{\lceil n/2 \rceil}}{2}$$

Мода

Частое значение

Усечённое среднее

$$\frac{x_k + \dots + x_{m-k+1}}{m - 2k + 2}$$

- + весовые схемы
- + сглаживание

$$\text{mid-range}(X) = \frac{x_1 + x_m}{2}$$

Описательные статистики – характерные элементы

Минимум

$$x_1$$

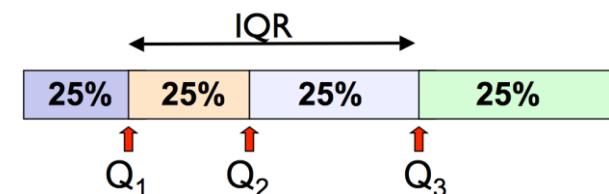
Максимум

$$x_m$$

Квантиль – значение, которое с.в. не превышает с заданной вероятностью

Квартили

$$q_{0.75}(X), q_{0.5}(X), q_{0.25}(X)$$



Децили

$$q_{0.1}(X), q_{0.2}(X), \dots, q_{0.8}(X), q_{0.9}(X)$$

Процентили

$$q_{1\%}(X), q_{2\%}(X), \dots, q_{98\%}(X), q_{99\%}(X)$$

Описательные статистики – разброс значений

Среднее линейное (абсолютное) отклонение
Mean Absolute Deviation

$$\frac{1}{m} \sum_{i=1}^m |x_i - m(X)|$$

m(X) – любая формализация среднего

Среднеквадратическое отклонение
Mean Squared Error (MSE) / Mean Squared Deviation (MSD)

$$\sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - m(X))^2}$$

Описательные статистики – абсолютные вариации

Чаще: стандартное отклонение

$$\text{std}(X) = \sqrt{\frac{\sum_{i=1}^m (x_i - \text{mean}(X))^2}{m-1}}$$

Дисперсия (рассеяние, разброс)

$$\text{var}(X) = \text{std}^2(X)$$

Размах

$$\text{range}(X) = x_m - x_1$$

Median Absolute Deviation (MAD)

$$\begin{aligned}\text{MAD}(X) &= \\ &= \text{median}(\{\text{median}(X) - x_i\}_{i=1}^m)\end{aligned}$$

Среднее квартильное расстояние

Интерквартильный размах

$$q_{0.75}(X) - q_{0.25}(X)$$

Описательные статистики – абсолютные вариации

Совет:

$$m_2(\{|x_i - m_1(X)|\}_{i=1}^m)$$

m_2, m_1 – любые формализации среднего

**Есть фундаментальный подход к оценке среднего,
а вариация описывается с помощью него**

Описательные статистики – относительные вариации

абсолютная вариация / среднее

Коэффициент вариации

Coefficient of variation

$$\frac{\text{std}(X)}{\text{mean}(X)}$$

Индекс дисперсии

Index of dispersion

$$\frac{\text{std}^2(X)}{\text{mean}(X)}$$

Относительный размах вариации (коэффициент осцилляции)

$$\frac{\text{range}(X)}{\text{mean}(X)}$$

Описательные статистики – другие...

Коэффициент асимметрии

оценка

$$\frac{\mathbf{E}[(X - \mathbf{E}X)^3]}{\mathbf{D}[X]^{3/2}}$$

Коэффициент эксцесса (островершинности)

оценка

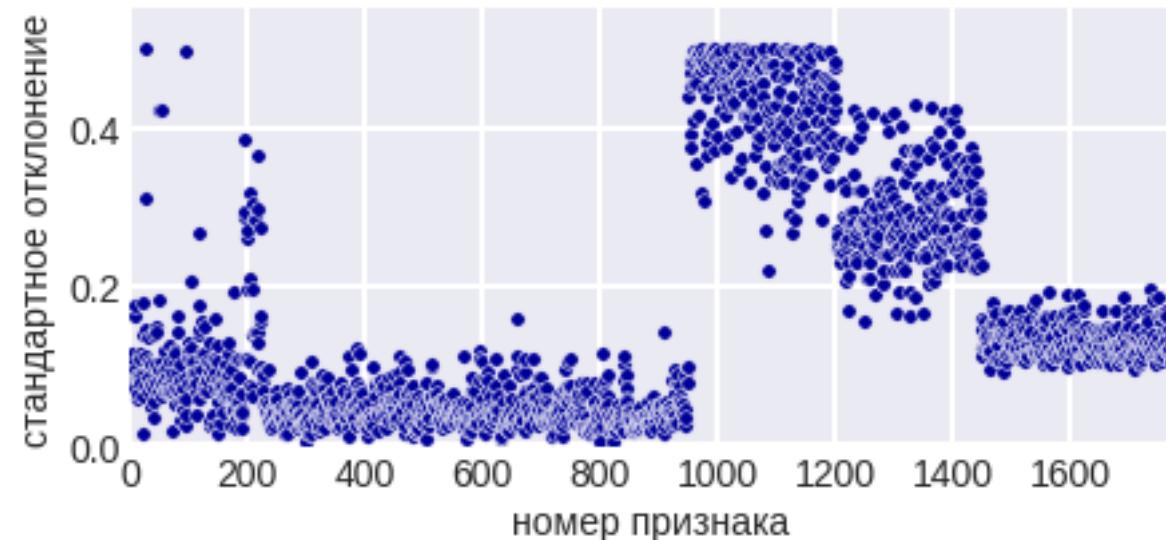
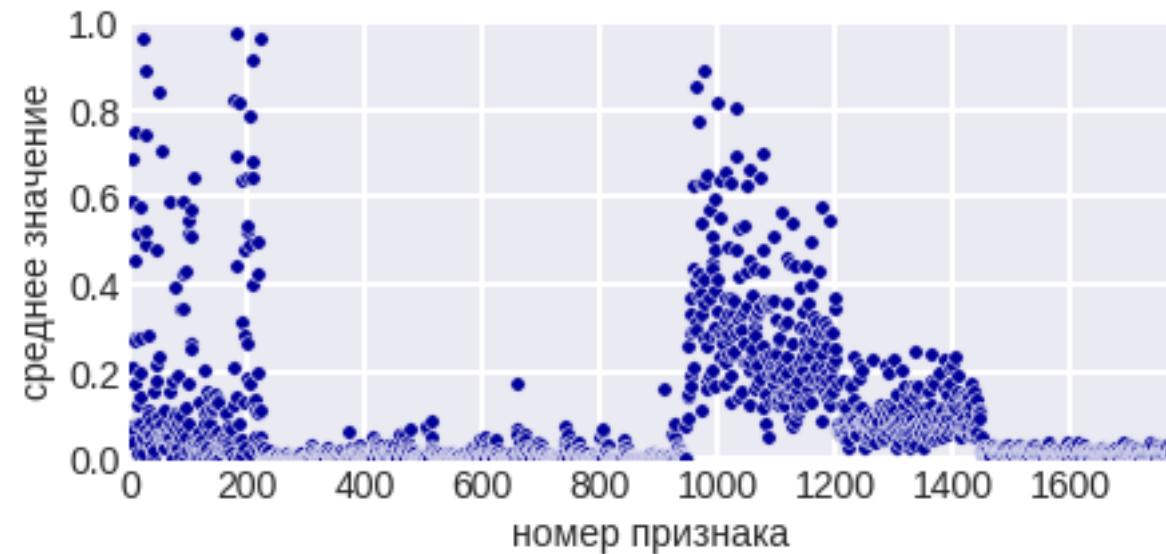
$$\frac{\mathbf{E}[(X - \mathbf{E}X)^4]}{\mathbf{D}[X]^2} - 3$$

Стандартная ошибка среднего

$$\frac{\text{std}(X)}{\sqrt{m}}$$

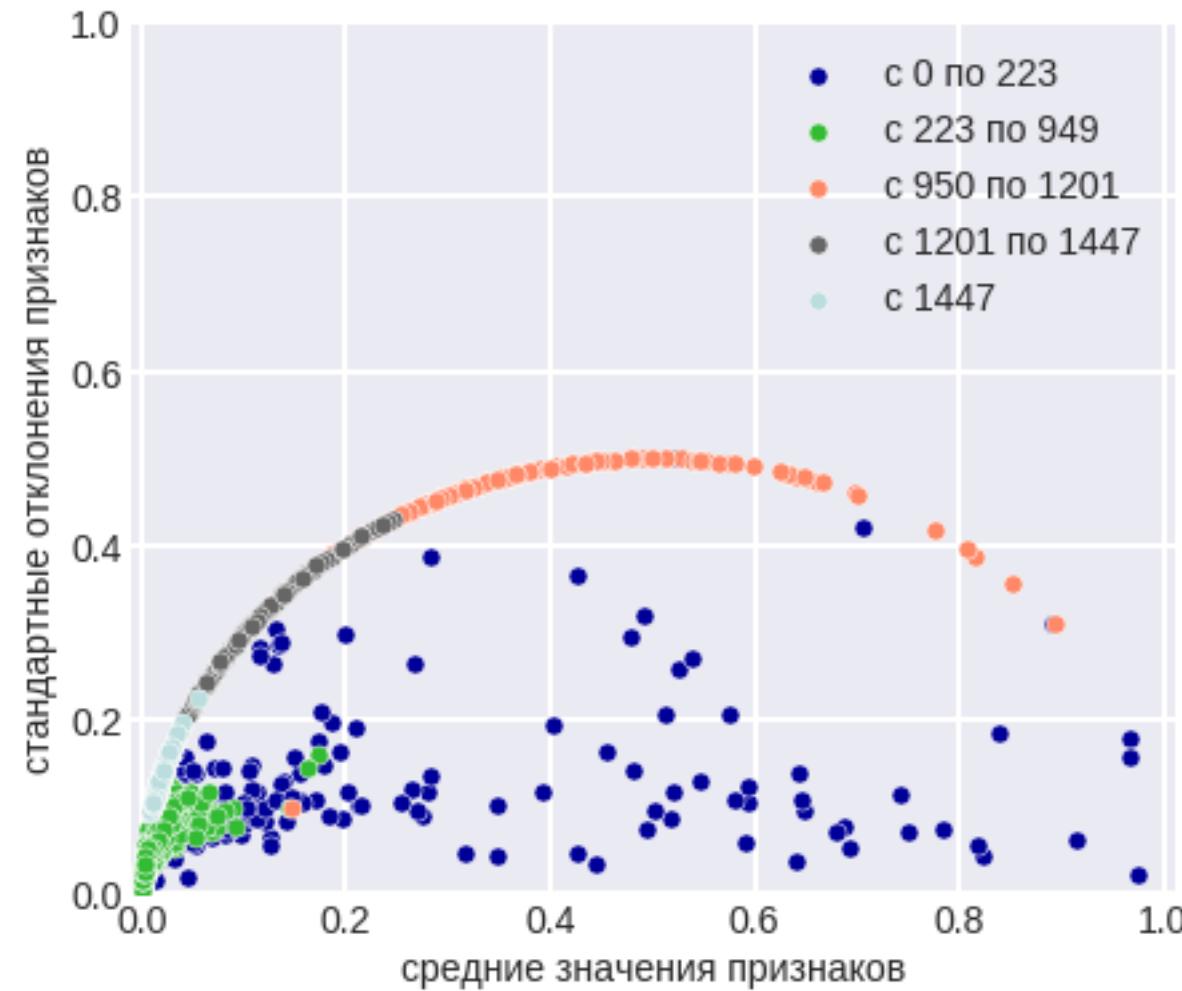
ДЗ Понятные и простые иллюстрации + ещё описательные статистики

ЗАДАЧА BIOLOGICAL RESPONSE



Чётко видны группы

Фантастика: дугообразная зависимость у трёх групп признаков!



ВОПРОС: Какие это признаки?

ОТВЕТ: это были бинарные признаки!

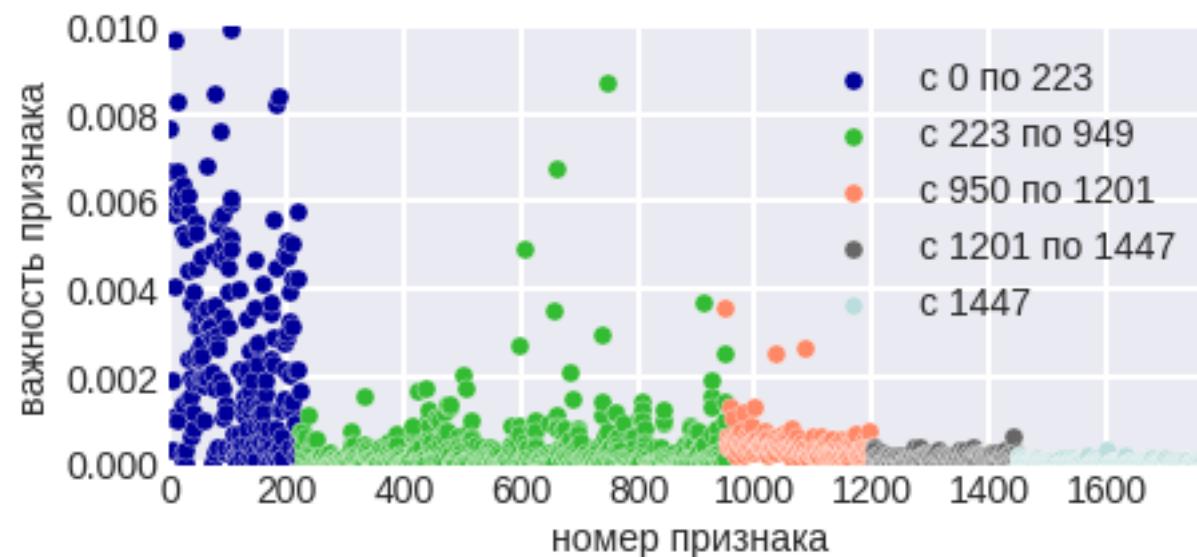
У них std зависит от mean (поскольку $x_i^2 = x_i$)!

[0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0]

$$\text{mean}(\{x_i\}_{i=1}^m) = \frac{1}{n} \sum_{l=1}^n x_i \equiv p$$

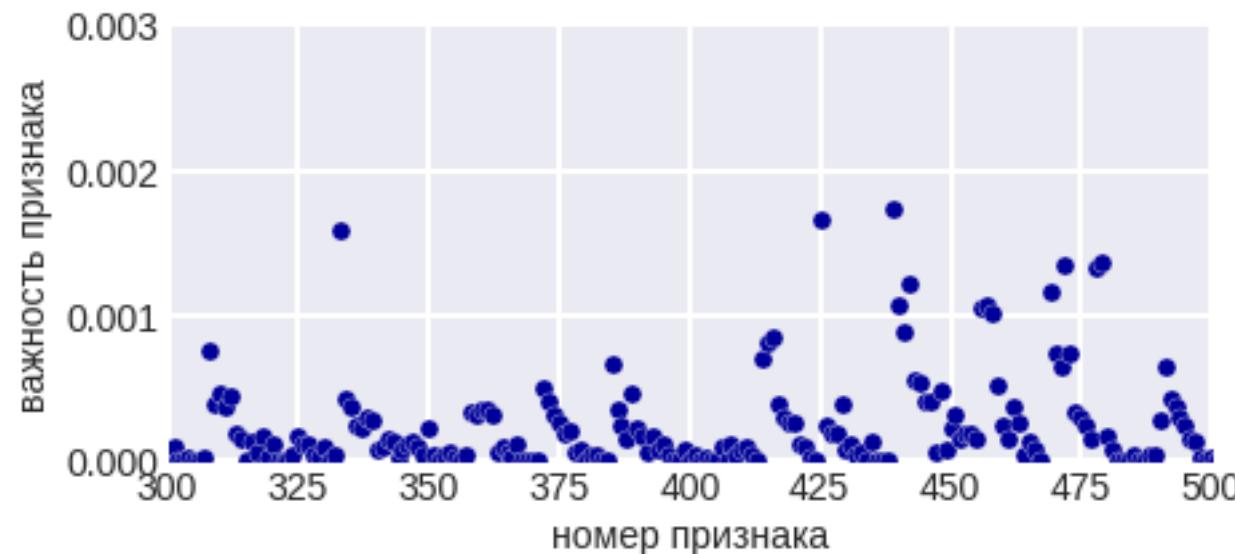
$$\begin{aligned}\text{std}(\{x_i\}_{i=1}^m) &= \sqrt{\frac{1}{m} \sum_{i=1}^m \left(x_i - \frac{1}{m} \sum_{l=1}^m x_i \right)^2} = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - p)^2} = \\ &= \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i^2 - 2px_i + p^2)} = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - 2px_i + p^2)} = \\ &= \sqrt{\frac{1-2p}{m} \sum_{i=1}^m x_i + p^2} = \sqrt{(1-2p)p + p^2} = \sqrt{p-p^2} = \sqrt{p(1-p)}\end{aligned}$$

Важности признаков с точки зрения RF.



**Потом: целые группы признаков можно удалять
без существенной потери качества**

Увеличение картинки

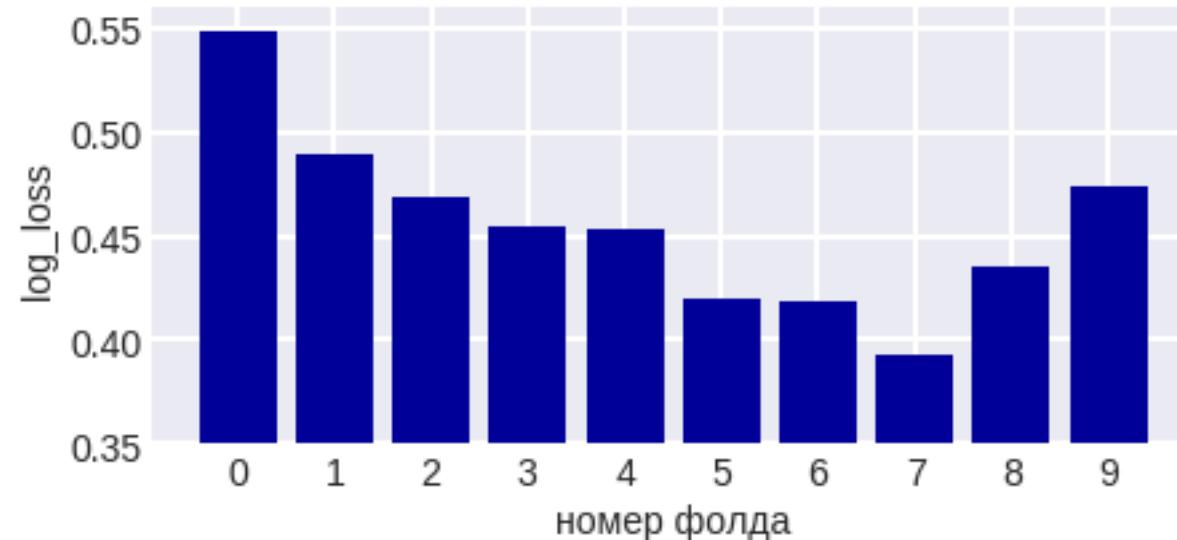


Есть подгруппы признаков!

Меняйте масштаб!

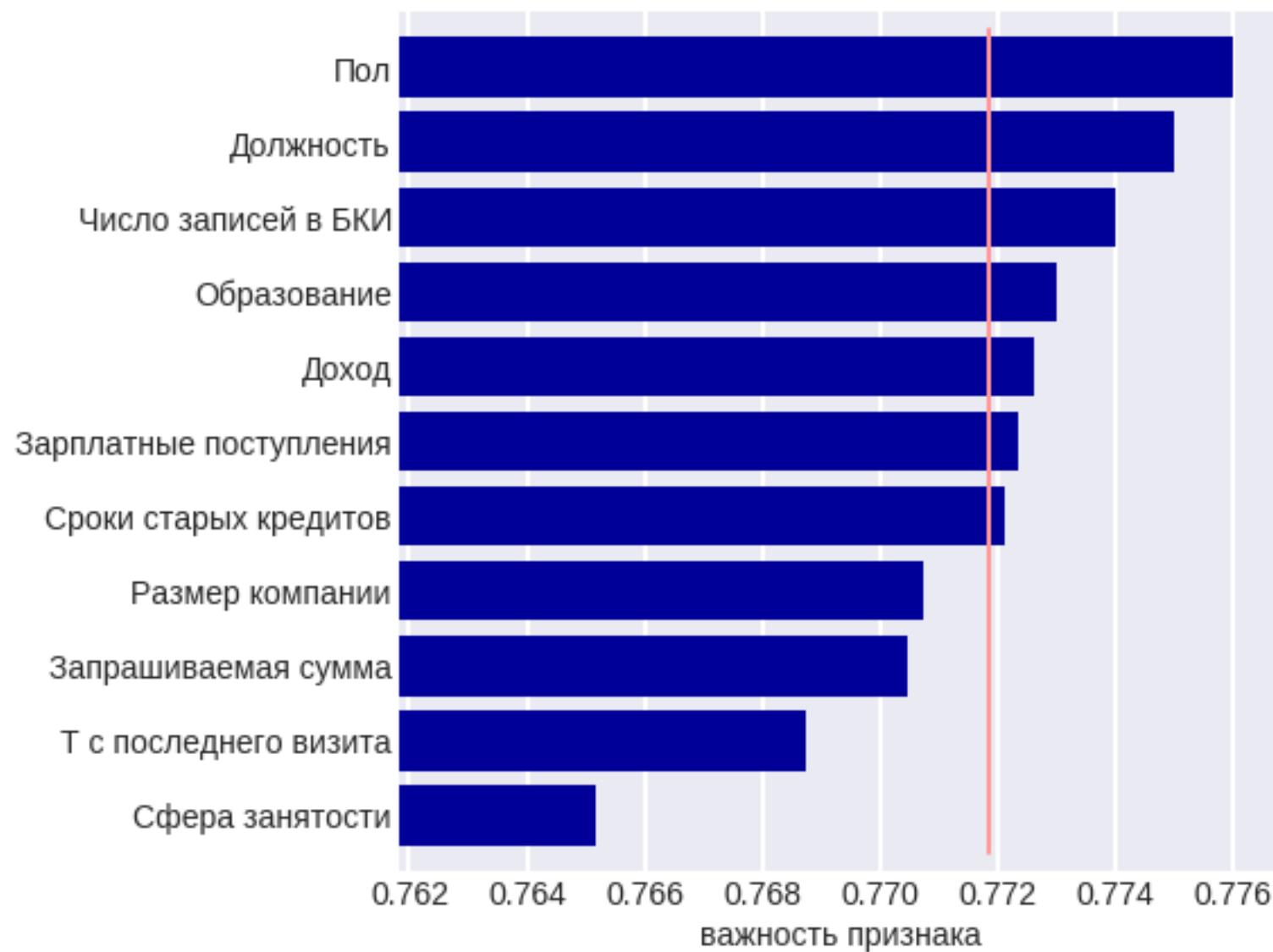
Аналогично: исследование сложности «классификации» объектов

Исследование частей выборки (фолдов)



**Подозрительная унимодальная зависимость!
Что значит?**

Как правильно показывать важности признаков



Сортировка, среднее значение, вертикальная ориентация

Правило столбцовых диаграмм



Правило:

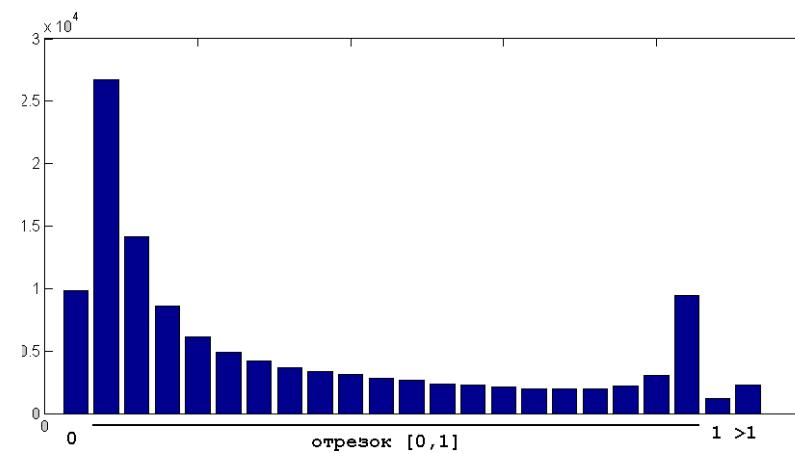
- **упорядочивать по убыванию/возрастанию показателя
(а не по алфавиту)**
- **Дать ориентир – что хорошо / что плохо**
- **Правильная ориентация делает визуализацию понятнее**

Что часто делается в начале задачи

Задача «Give Me Some Credit»

Статистика признаков

признак	%	Age	Доход	#90	#	#60	# в сем
значения	[0, 1] есть дроби!	0, 1, 21-109	целые	0-17, 96, 98	0-26, 32, 54	0-9, 96, 98	0-10, 13, 20
# уникальных значений	84500	86	11866	19	26	12	13
неизвестных значений			19831				
AUC	0.7815	0.6329	0.5554	0.6613	0.5432	0.6247	0.5499



Смотрим на сами признаки

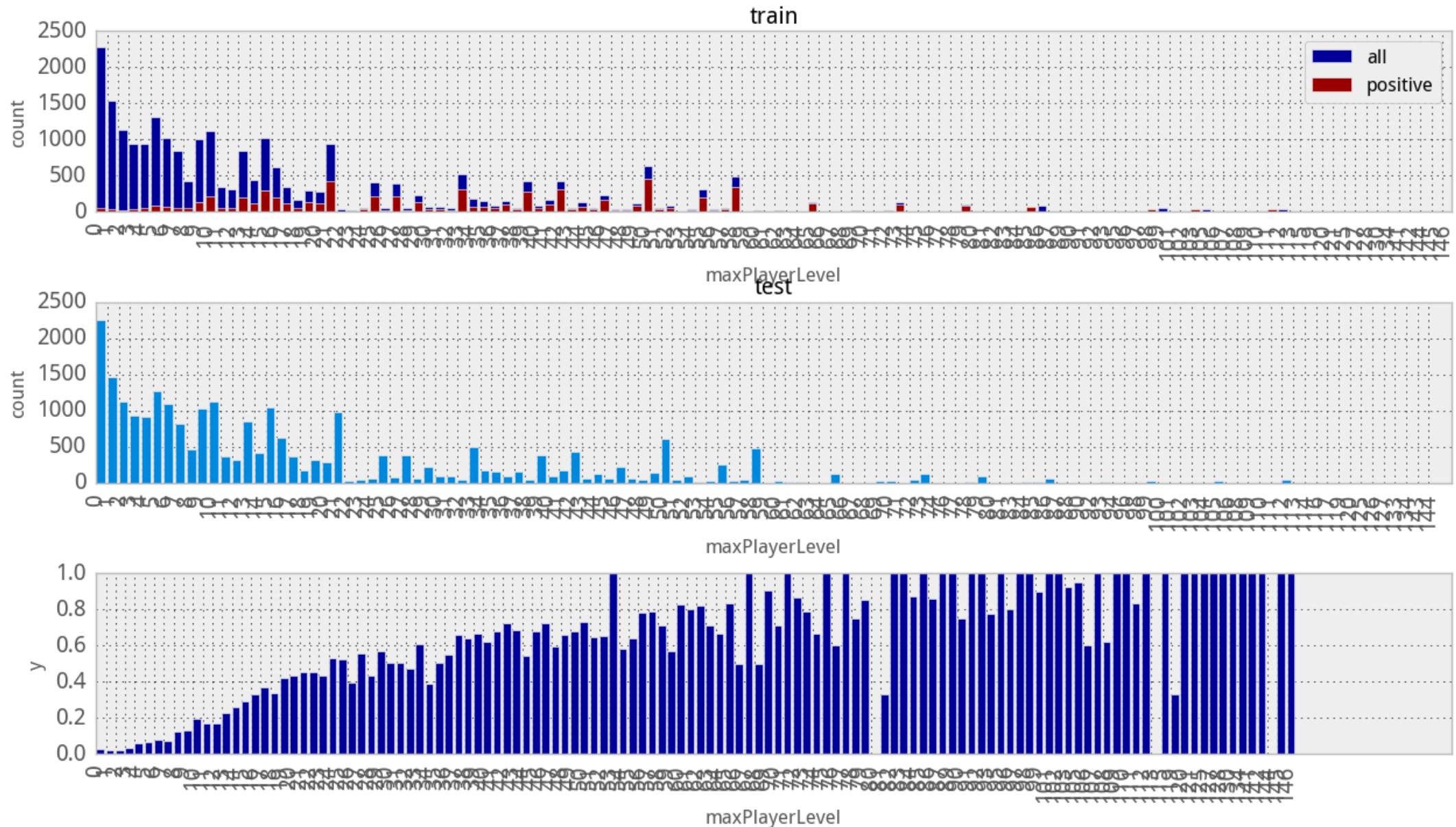
```
for name in data.columns:  
    if data[name].nunique() < 8:  
        u = data[name].unique()  
    else:  
        u = data[name].unique()[:8]  
    if type(data[name].tolist()[0]) is str:  
        print ('%25s %10d %10s %10s %s' % (name, data2[name].nunique(), '', 'str', str(u)))  
    elif type(data2[name].tolist()[0]) is pd.tslib.Timestamp:  
        print ('%25s %10d %10s %10s %s' % (name, data2[name].nunique(), '', 'time', ''))  
    else:  
        print ('%25s %10d %10.2f %10.2f %s' % (name, data2[name].nunique(), data2[name].mean(),  
                                                data2[name].std(), str(u)))
```

Класс	4	2.20	0.97	[1 2 3 4]
Номер	8404	7442.45	269.63	[5001 5002 ...]
Вес, т	124	38.27	7.30	[41.1 44.4 ...]
Начало	8404		time	
Количество, шт	45	63.78	5.13	[66. 61. ...]

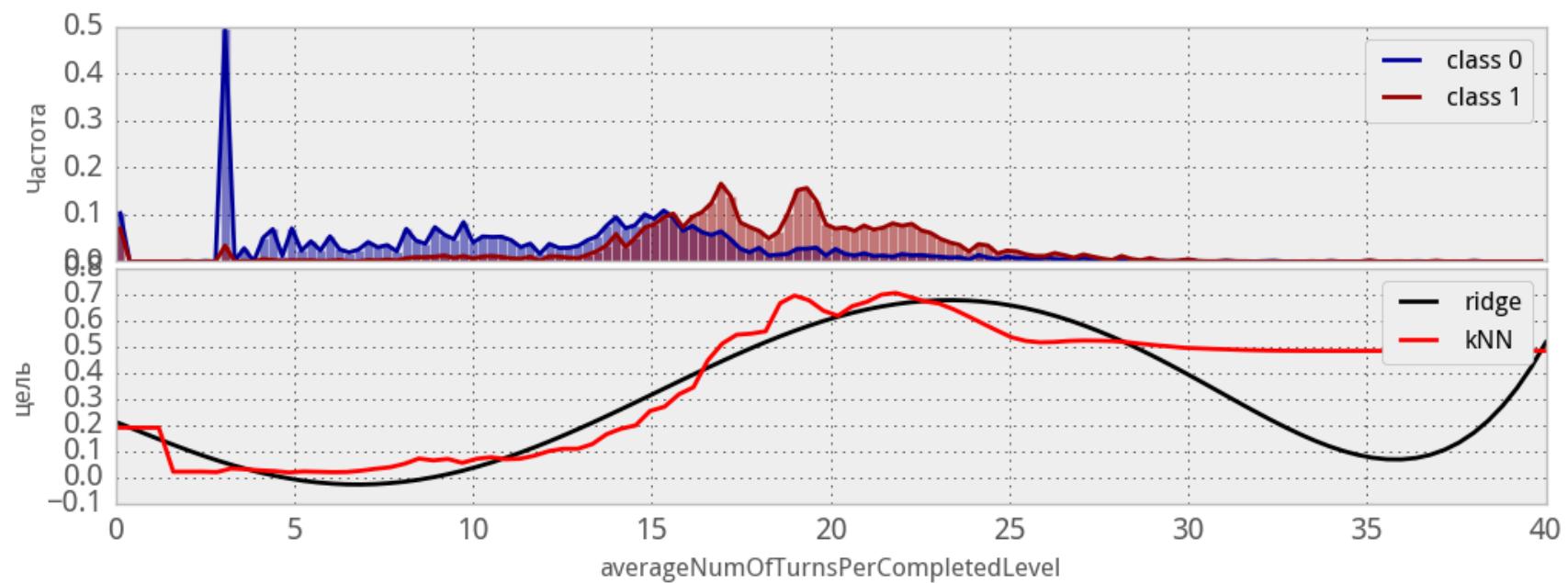
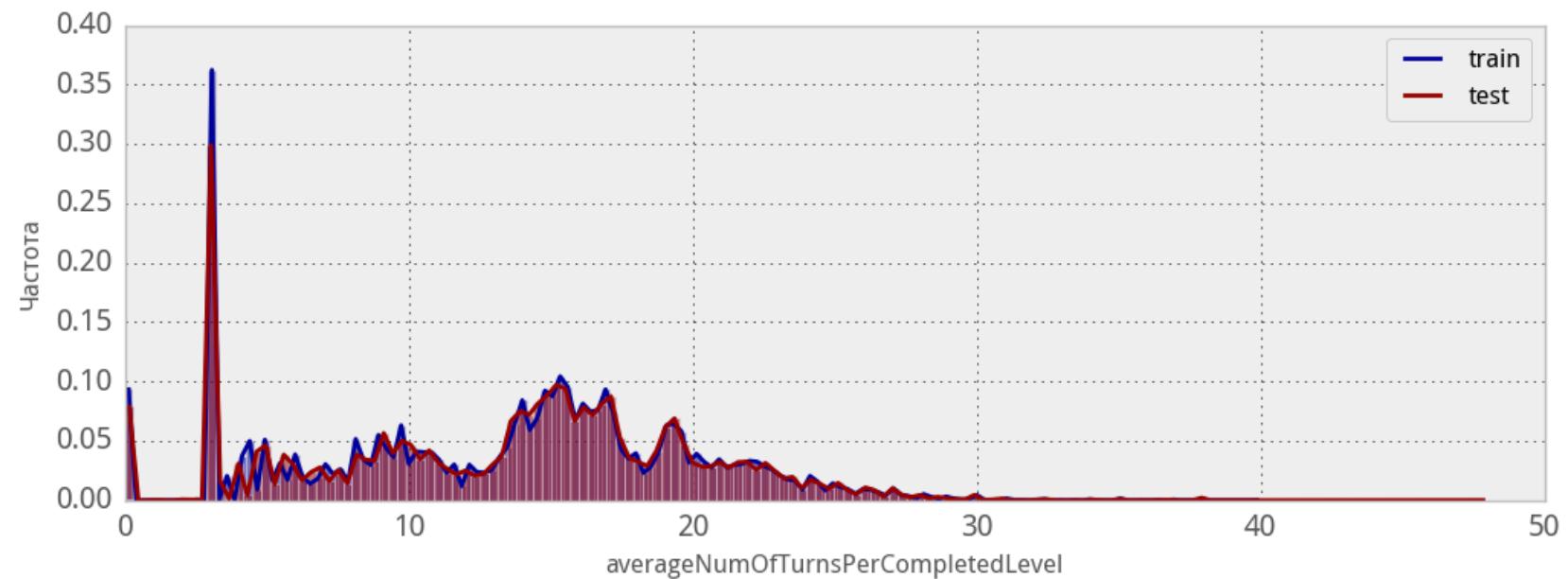
Что надо сразу выяснить про признак

- **распределение обучения / тест**
 - **распределение класс 0 / 1**
- **значения целевой переменной по категориям**

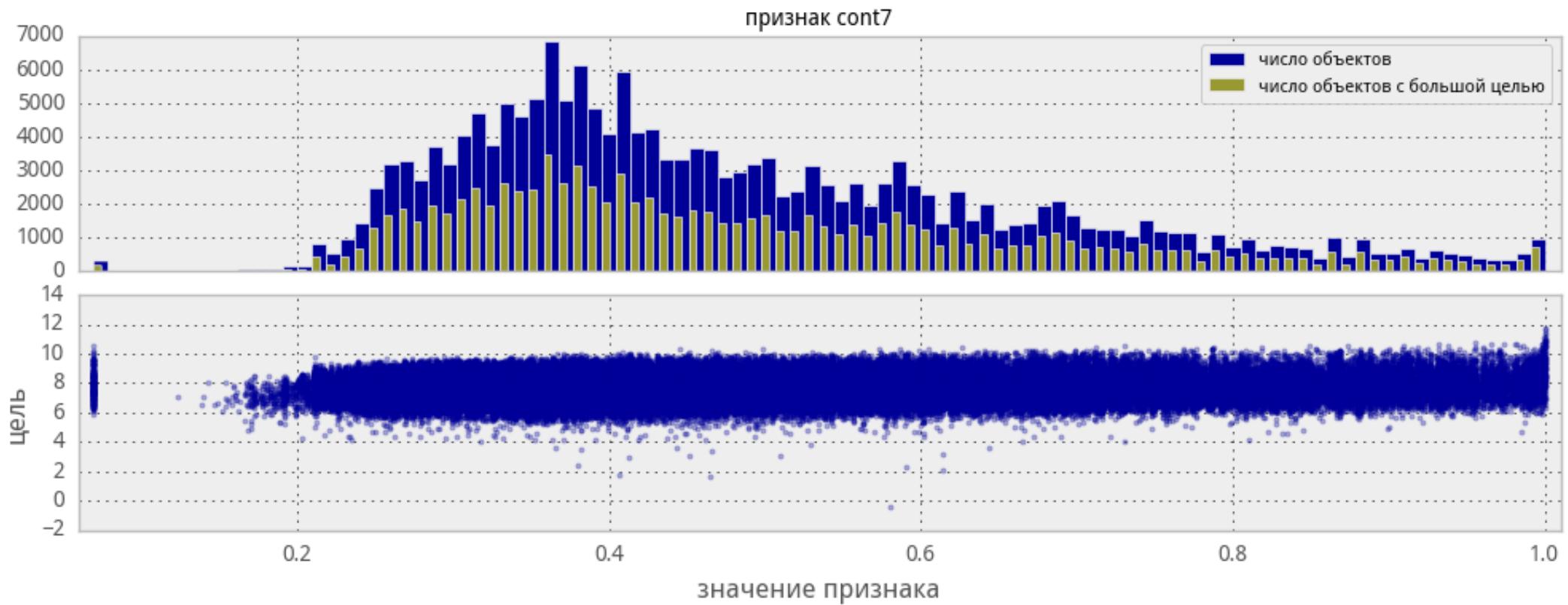
Что надо сразу выяснить про признак



Что надо сразу выяснить про признак



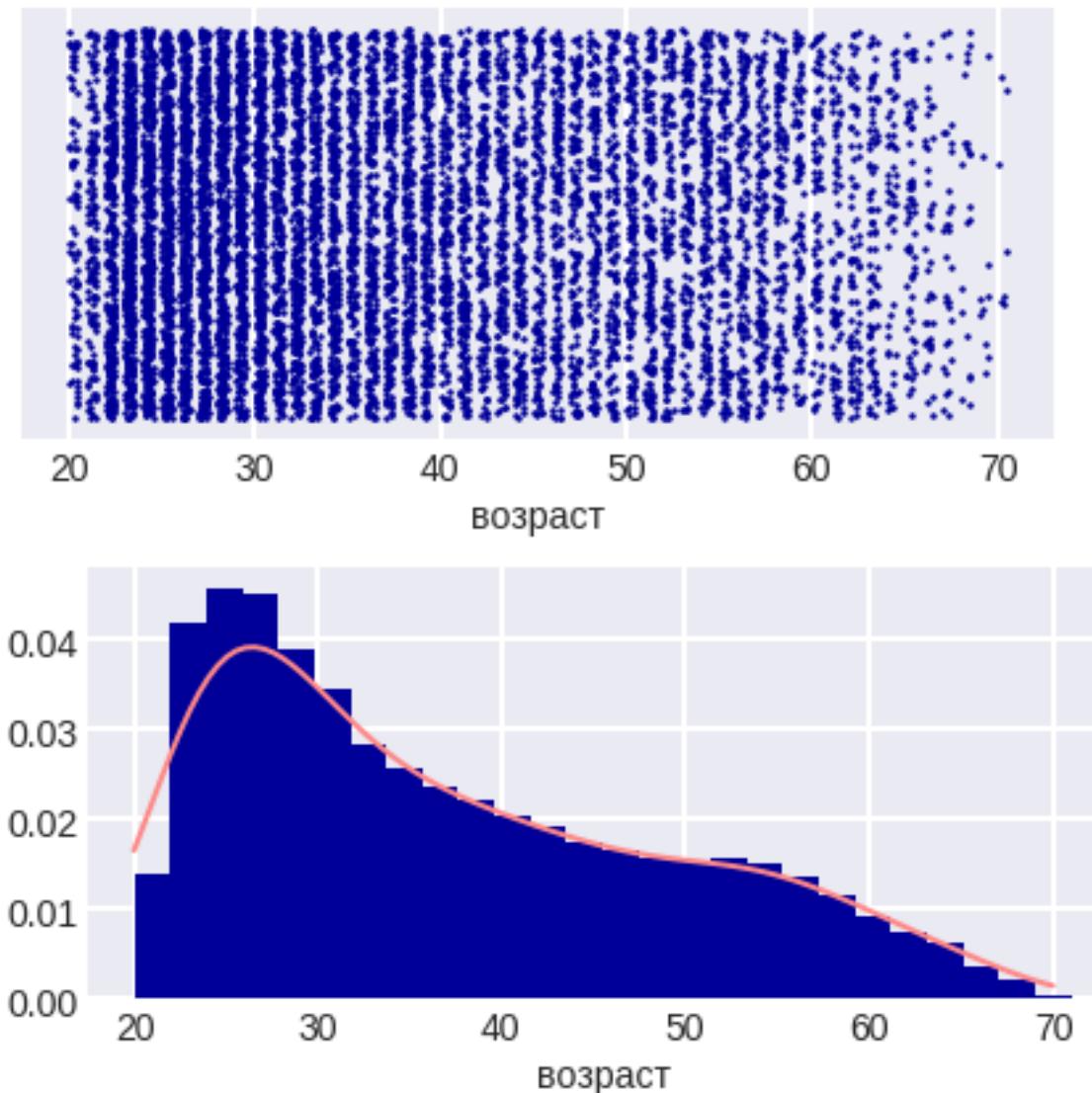
Что надо сразу выяснить про признак «AllState»



**Верху – гистограмма распределения по значениям признака
Отдельно по объектам с большим значением целевого признака**

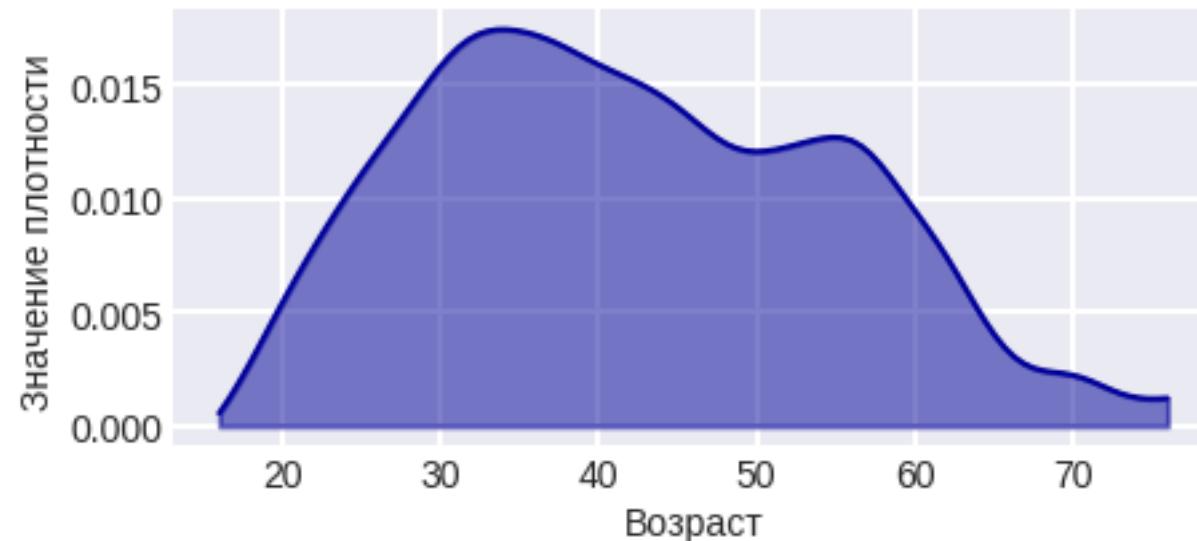
Внизу – диаграмма рассеивания «признак – цель»

Визуализация отдельных признаков



Недаром нужны гистограммы!

ЗАДАЧА «М-магазин»

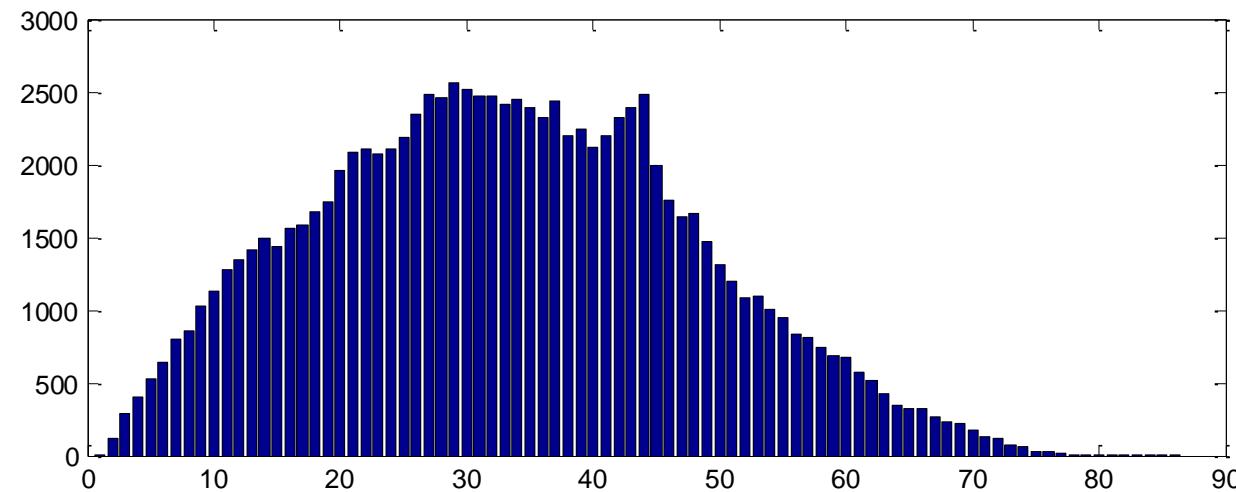


Распределение возраста покупателей

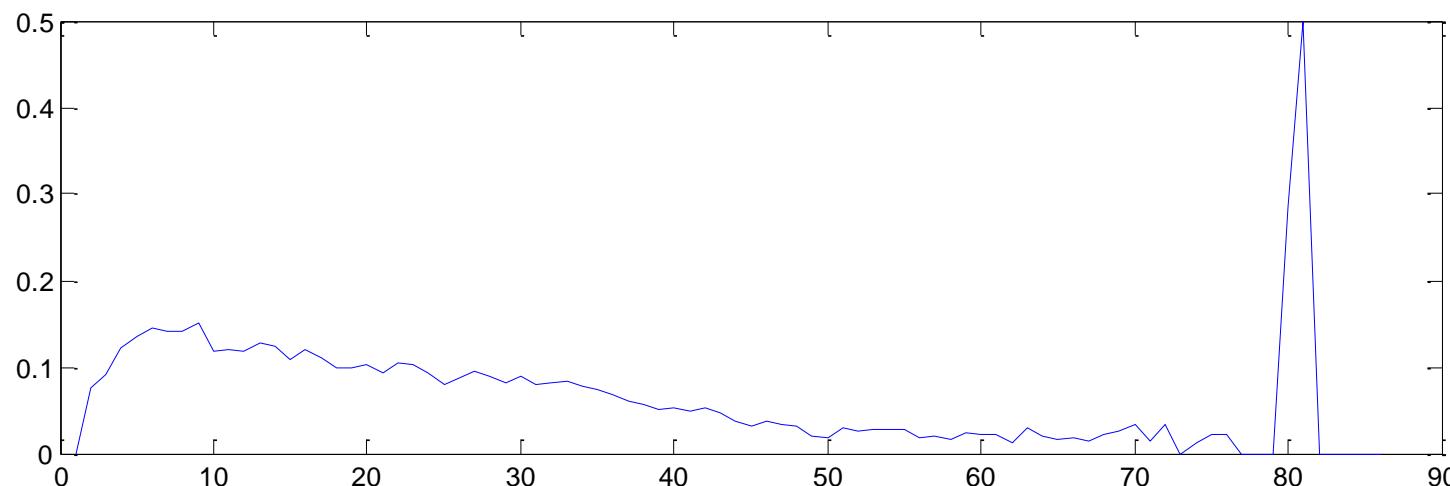
Так обычно выглядит распределение!

Почему два горба?

ЗАДАЧА «ТКС»

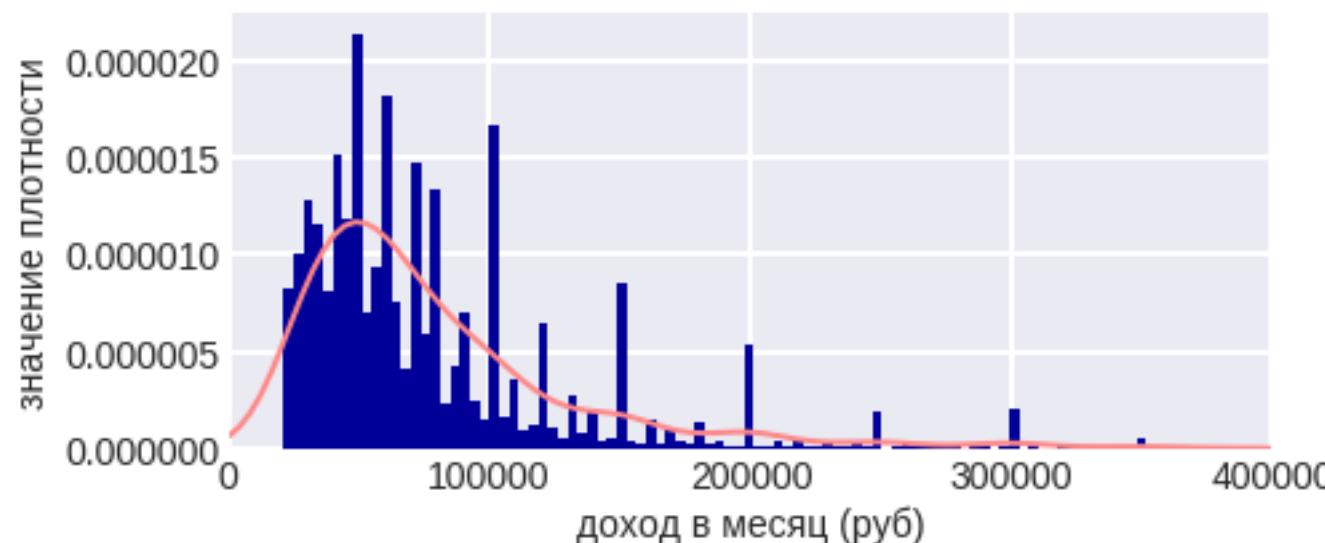
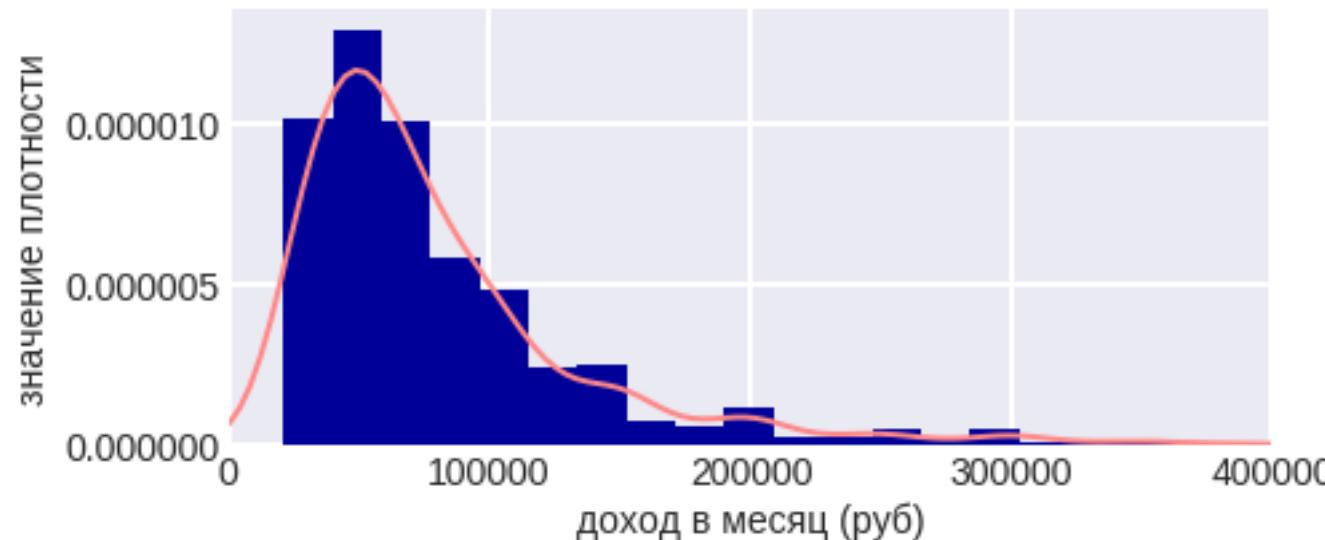


Распределение по возрасту
Что значит?



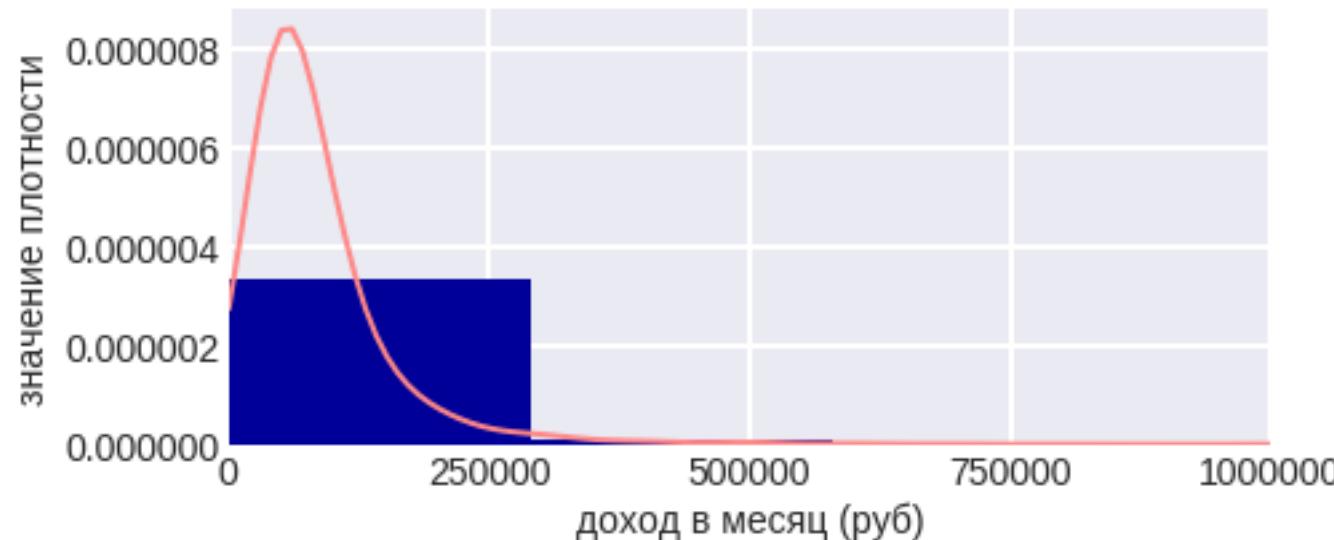
Отношение плотностей – есть явный выброс!

Проблемы визуализаторов – параметры по умолчанию



увеличили число бинов

Проблемы визуализаторов – выбросы

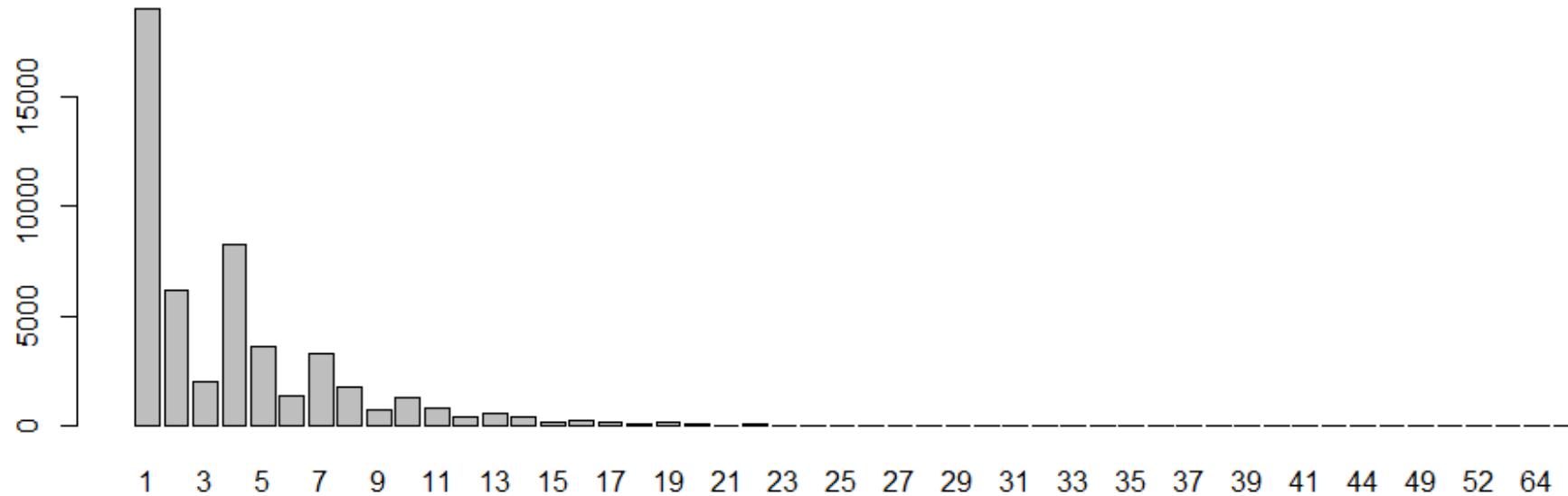


Что будет если не устранять выбросы...

```
def make_clips(data, name):  
    return (data[name].clip(lower=data[name].quantile(0.01),  
                           upper=data[name].quantile(0.99)).values)
```

Ещё раз о параметрах по умолчанию: «Liberty»

Что интересного в распределении целевого признака?
a transformed count of hazards or pre-existing damages

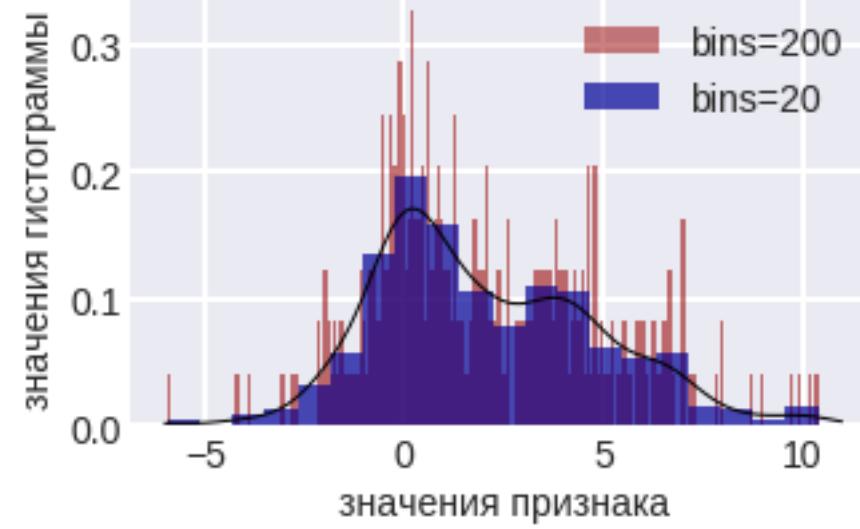
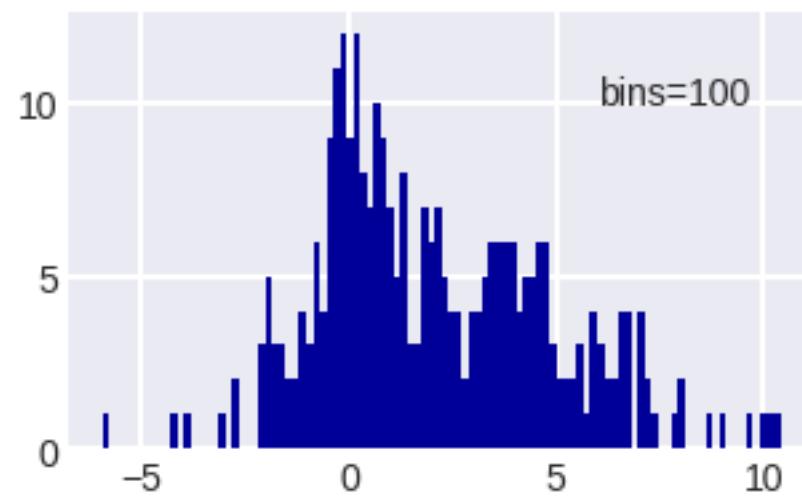
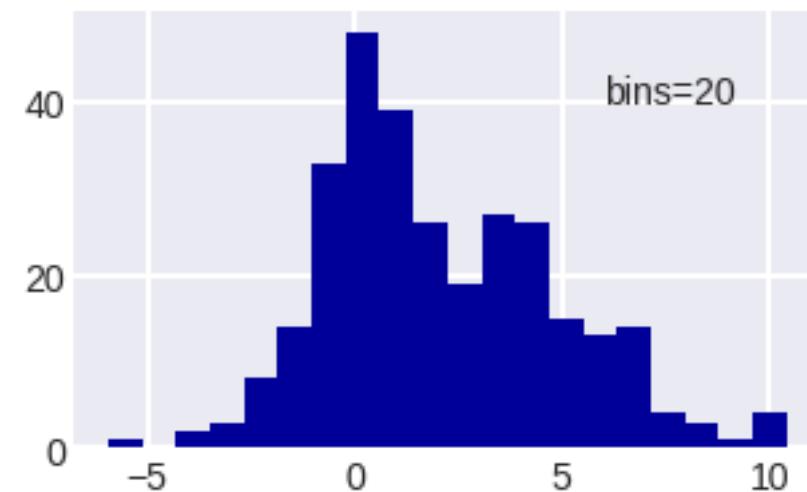
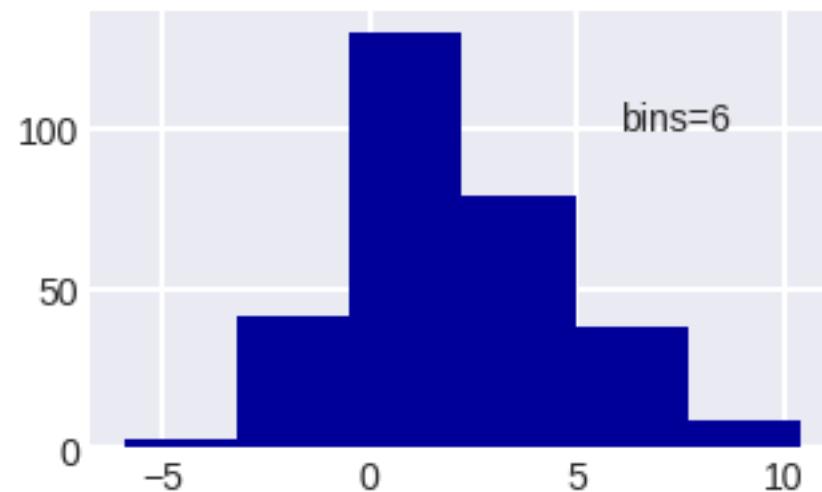


Ещё раз о параметрах по умолчанию: «Liberty»



Выбирать:
число бинов
ширина столбцов

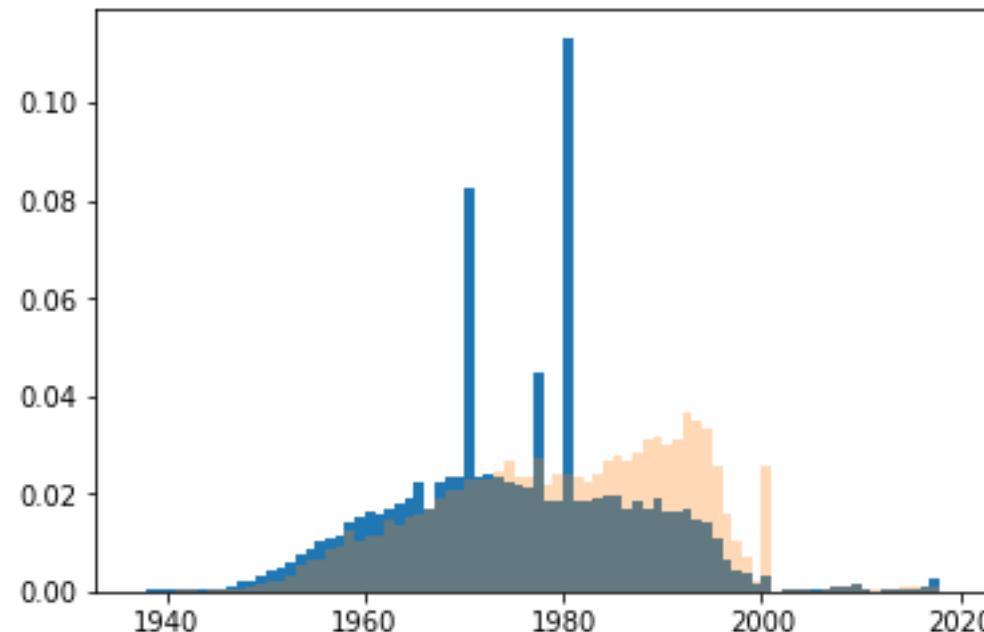
Построение гистограммы



Подбирайте число корзинок (бинов). Совет: можно совмещать!

Выводы о признаках

Распределения дат рождения пациентов (по полу)



Когда смотрим частые значения

1980-01-01	4850
1970-01-01	3013
1977-07-07	1321
2000-06-07	447
2017-04-01	155
2000-01-01	127
2009-04-01	109

Выводы о признаках

**значения по умолчанию ⇒ точная дата неизвестна
при этом пол «Ж» ⇒ тоже неверно
Стоит ли доверять другой информации?**

Визуализация отдельных признаков

Приёмы

- взять подвыборку
- менять число бинов!
- самому выбирать бины!

Зачем

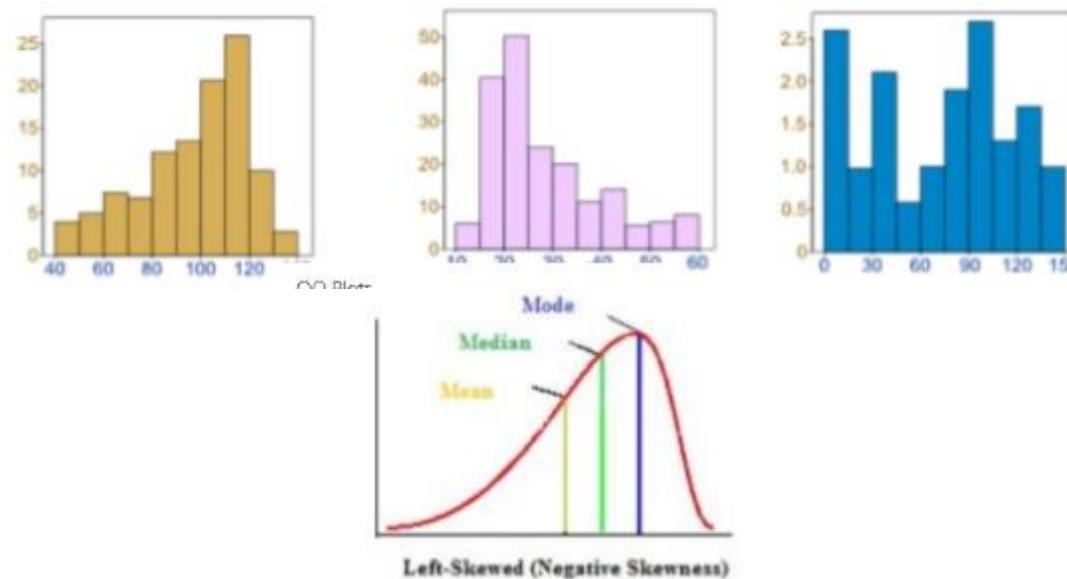
- логичность признака
- типичные значения
- области типичных значений
- преобразования признака

Сравнение:

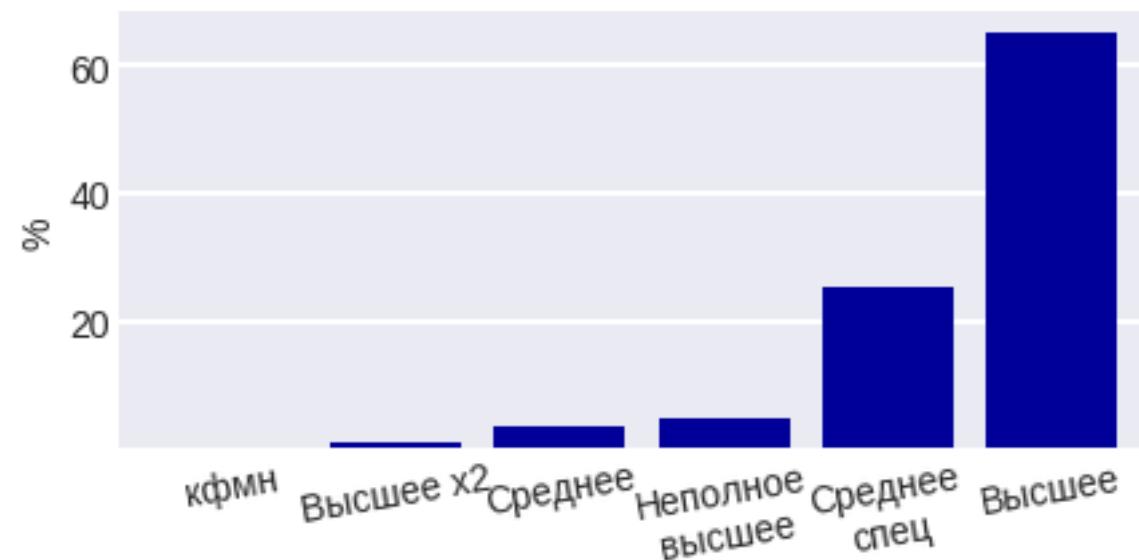
- при разных значениях целевого
- на обучении и контроле

Гистограммы NEW!!!

- быстро оценить форму распределения
 - придумать деформацию
 - зависит от организации бинов



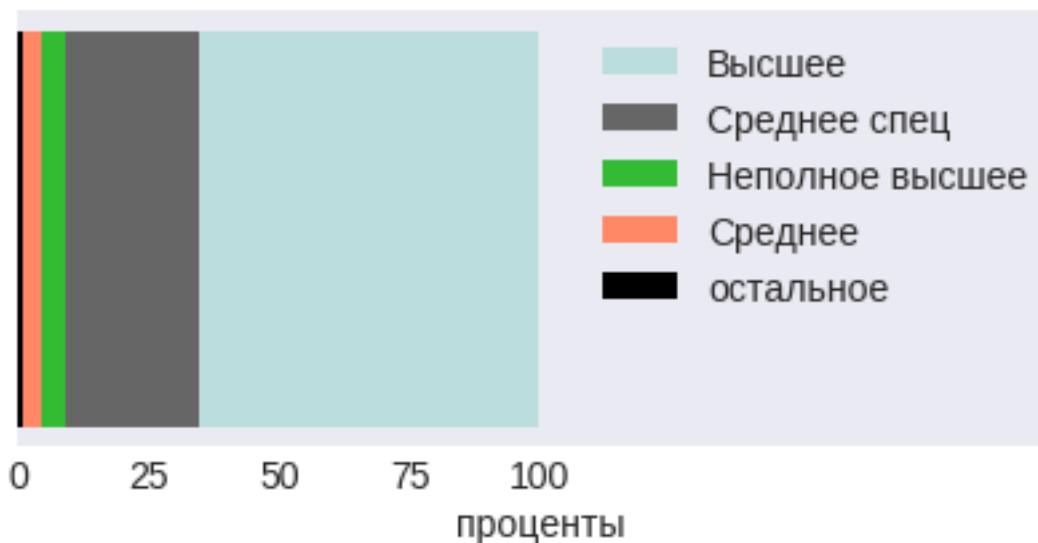
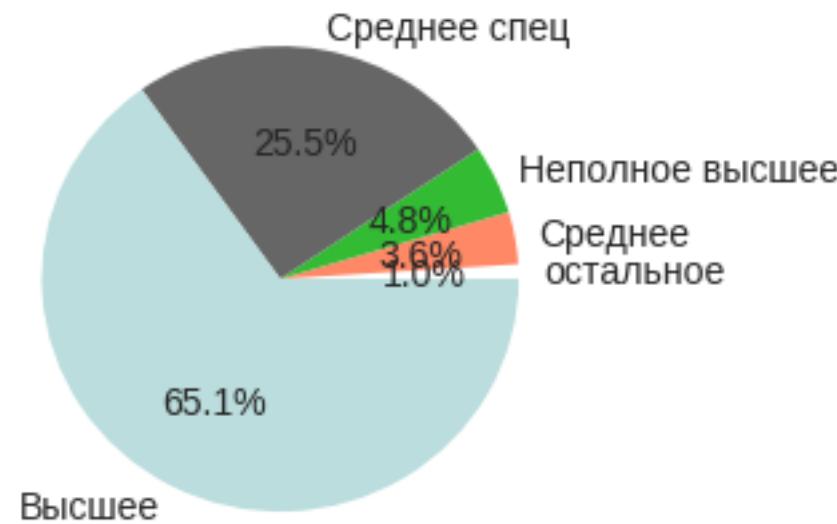
Визуализация категориальных признаков



**не видно мелкие категории
категорий может быть много**

Как быть?

Визуализация категориальных признаков



Визуализация категориальных признаков

Не использовать 3D-эффекты

Мелкие категории → «остальное»

Площадь всех категорий = 100%

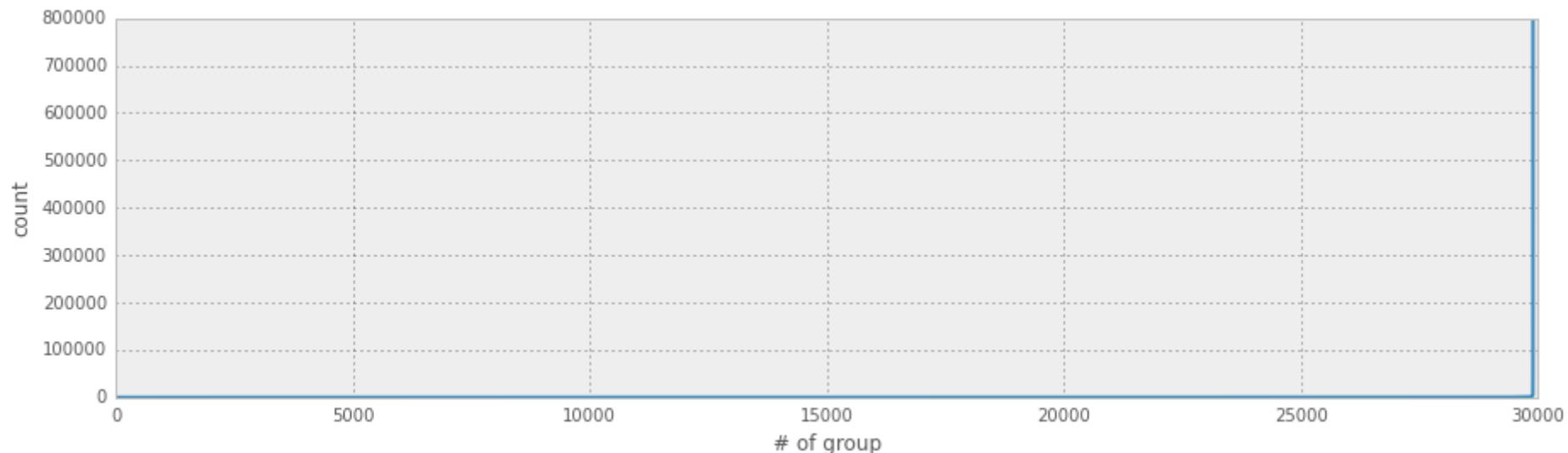
Диаграмма-пирог – не рекомендуется

Когда информации для визуализации мало – таблицы!

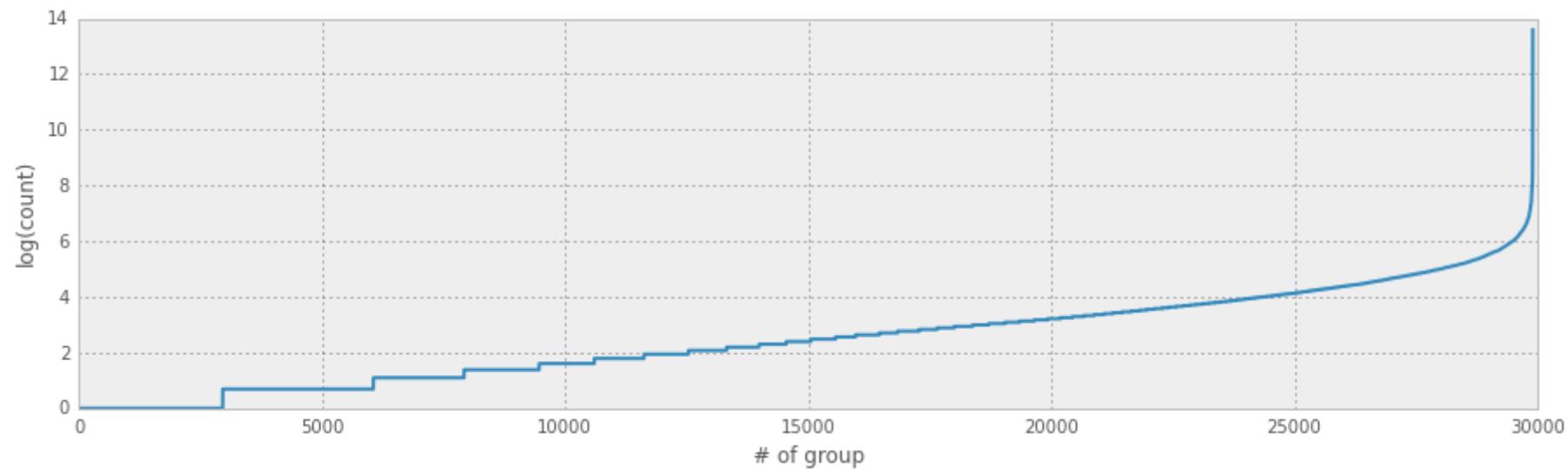
Образование	%
Высшее	65.1
Среднее спец	25.5
Неполное высшее	4.8
Среднее	3.6
Высшее x2	0.8
кфмн	0.2

Можно ещё логарифмировать...

Зачем ещё нужно логарифмирование



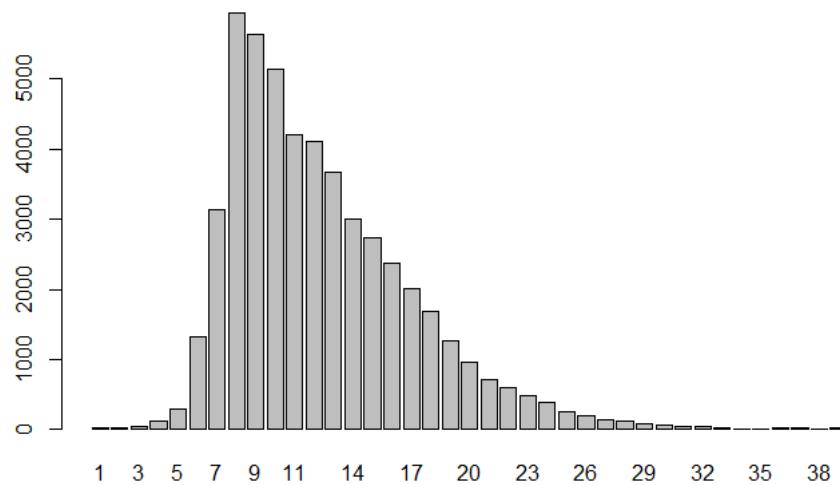
число представителей одной из ~30000 групп в выборке



логарифм этого числа

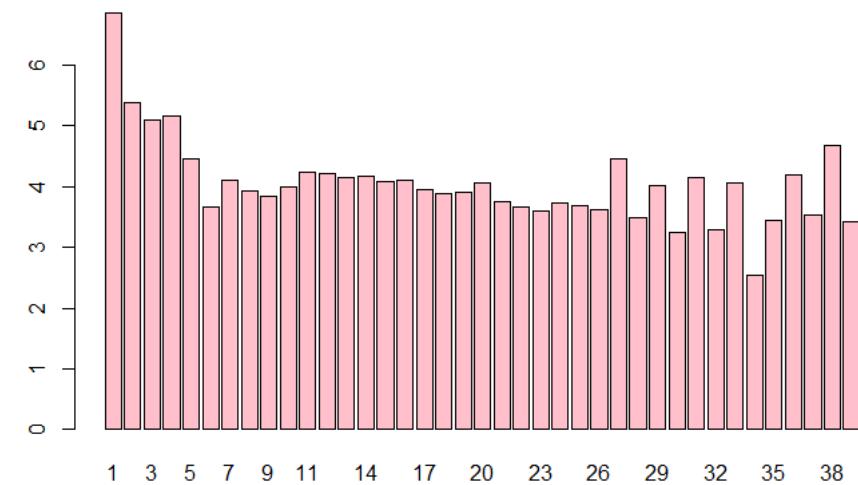
Распределения на признаках – природа признаков

**Задача «Liberty»: целочисленный признак –
вещественный или категориальный?**



```
barplot(table(train[,21]))
```

**Распределение значений
признака**



```
barplot(tapply(train$Hazard, train[,34], mean),  
       col='pink')
```

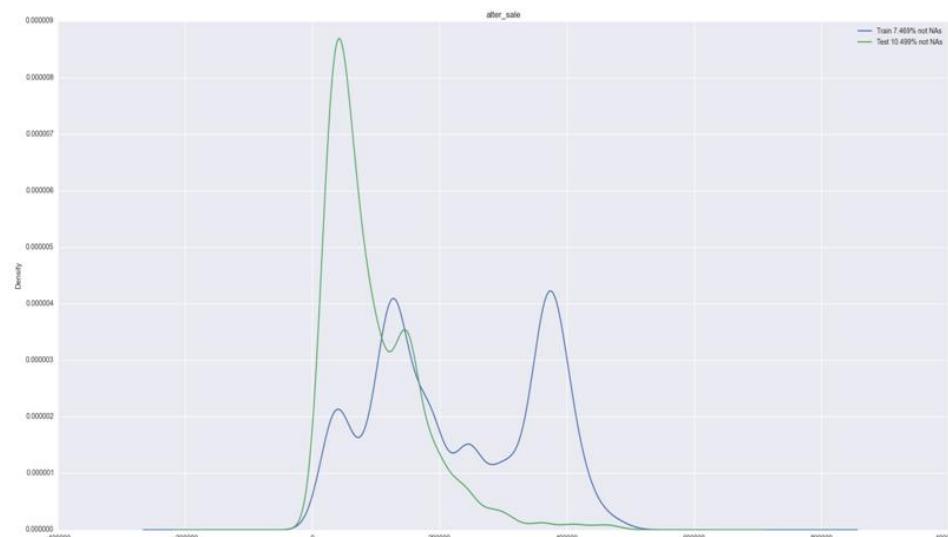
**Среднее цели на значениях
признака**

Категориальные признаки «AllState»

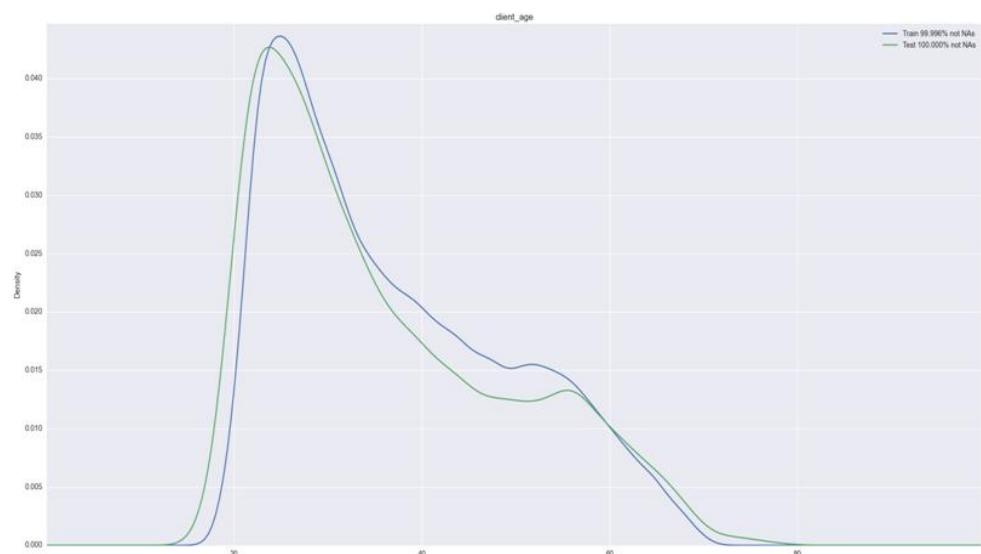
	mean	count		cat107	
	cat101		A	3259.510800	75
A	2454.139844	106721	B	19845.900000	2
B	1292.020000	3	C	2076.430704	213
C	2778.283638	16971	D	2636.230164	3225
D	2812.990306	17171	E	2871.429175	12521
E	4458.574286	7	F	3072.621189	47310
F	3560.151861	10139	G	3149.791915	28560
G	3450.680947	10944	H	3124.043153	23461
H	1320.720000	1	I	2913.988215	20066
I	4590.935254	6690	J	3084.531566	22405
J	4603.863790	7259	K	2946.549609	20236
K	3240.165000	2	L	3003.206170	6976
L	5321.419556	3173	M	3074.337929	2067
M	5540.292766	3669	N	3053.982033	797
N	2192.720000	1	O	2950.613520	125
O	6870.387172	2493	P	3138.672300	100
Q	7057.470264	2762	Q	2985.114143	140
R	8564.376594	138	R	3063.068000	5
S	8993.138439	173	S	5553.495000	2
U	15972.490000	1	U	3546.898438	32

Как распределение меняется при переходе к контролю

**смотреть как меняются распределения
обучение – контроль**



**Есть существенные
изменения**

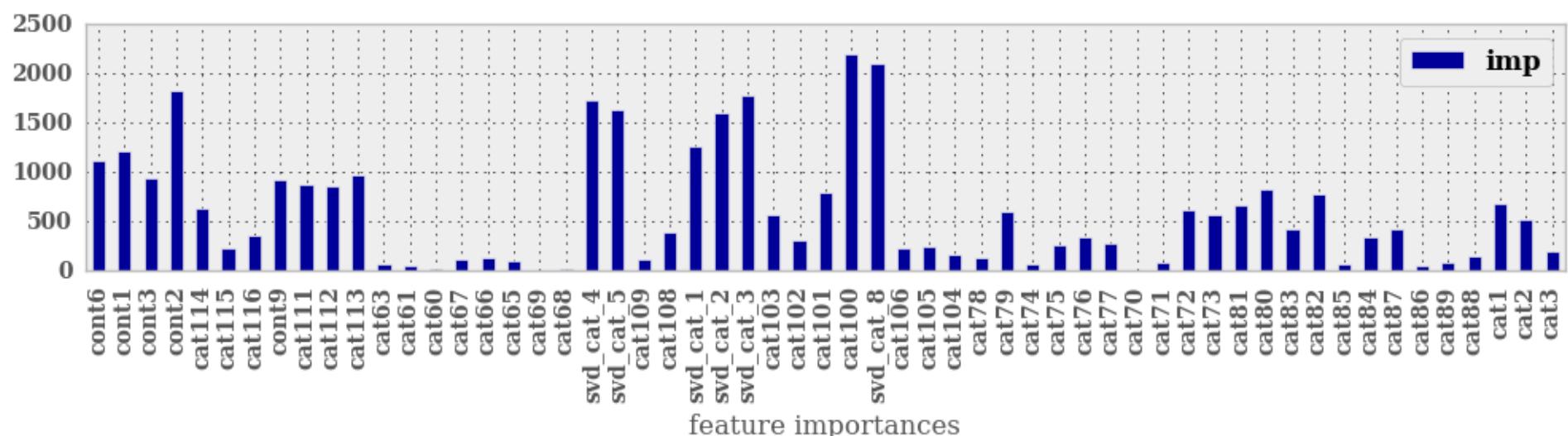
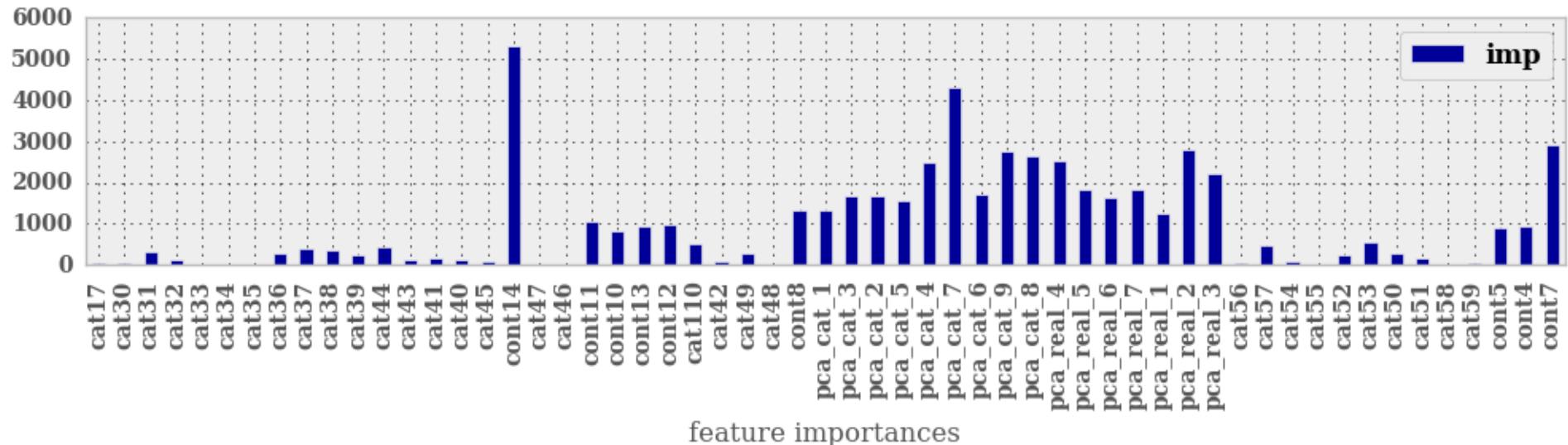


Нет изменений

История про о-трэвел и волшебный признак.

Важности признаков

Придумываем признаки и анализируем «AllState»

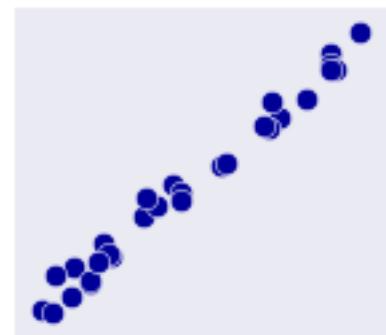


Визуализация пары признаков

**Самый распространённый способ –
диаграмма рассеивания («скатерплот»)**

А что на диаграмме рассеивания 2х признаков можно увидеть?

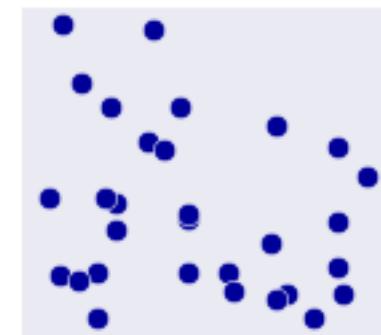
Что можно увидеть в данных («признак» – «признак»)



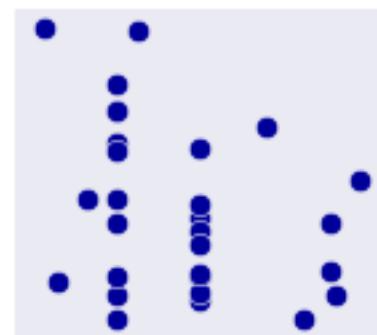
корреляция



зависимость



независимость



типичные значения



выбросы



кластеры

Что можно увидеть в данных («признак» – «признак»)
корреляцию

при правильном масштабе и небольшом шуме

зависимость признаков
при малом шуме и «достаточно равномерном» распределении

независимость признаков
часто это «ложное видение»

типичные значения
сложно при большом объёме данных

выбросы
при правильном масштабе

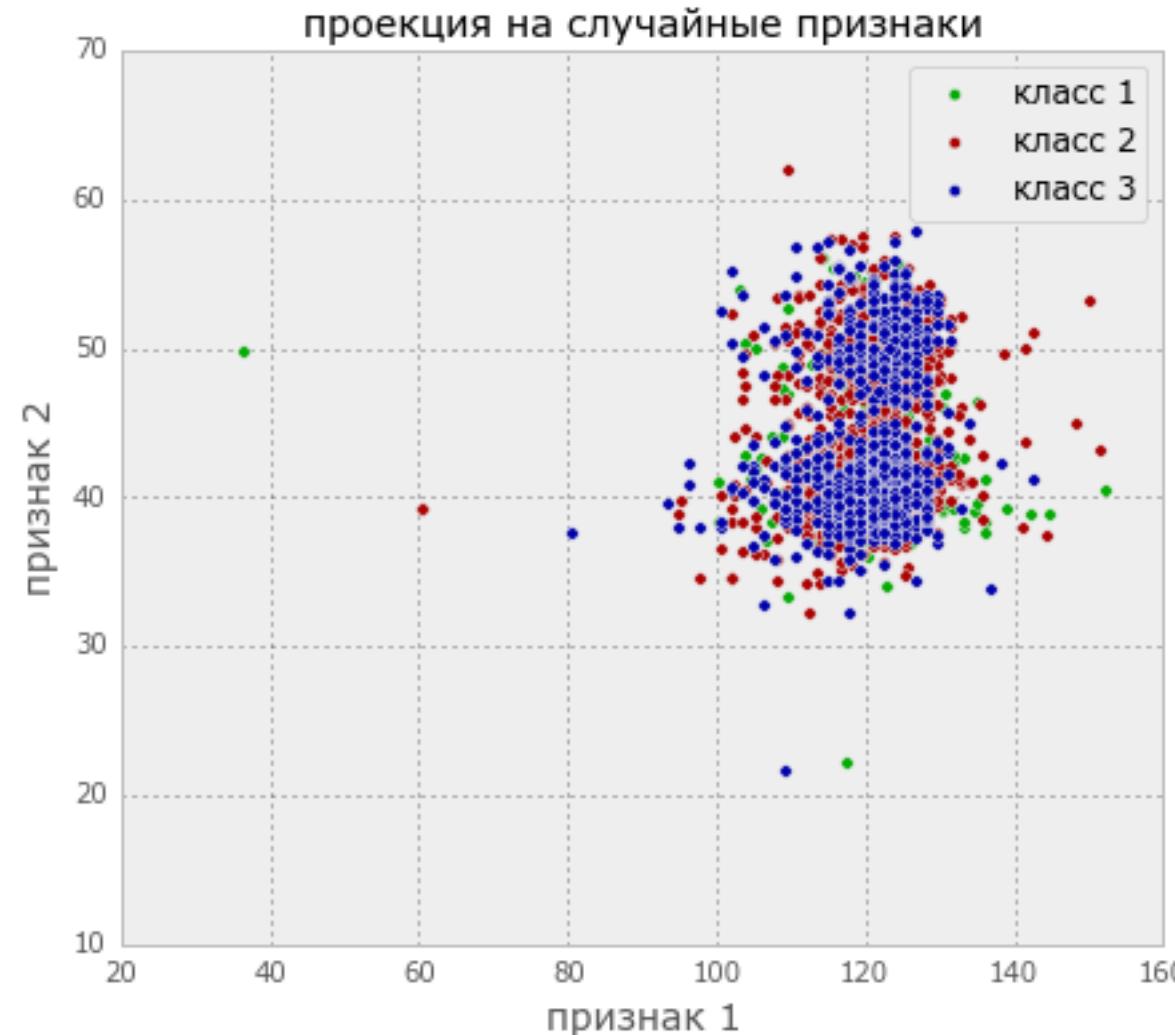
кластеры
при правильном масштабе

Смотрим на пары признаков

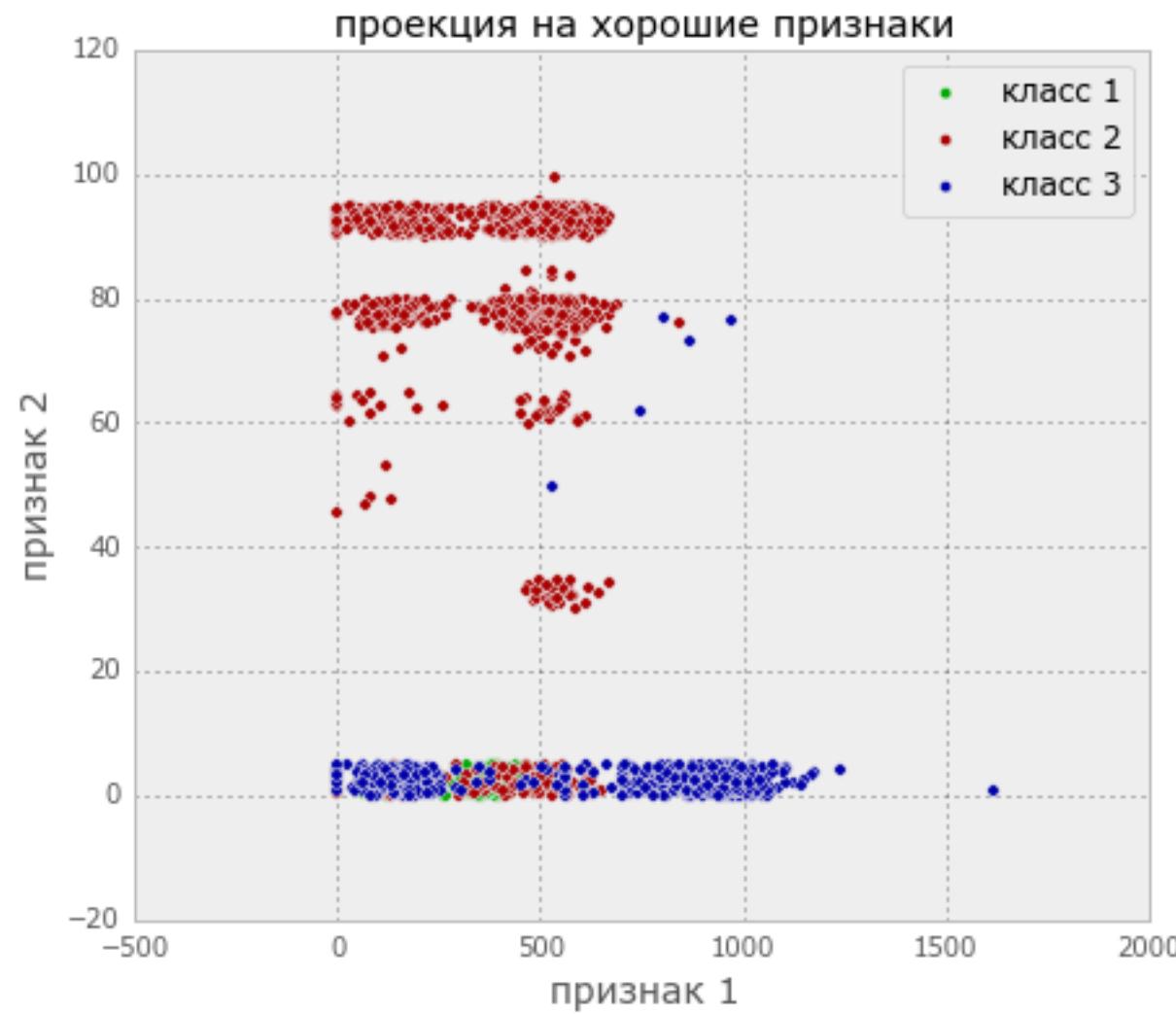
если есть время

признаков немного

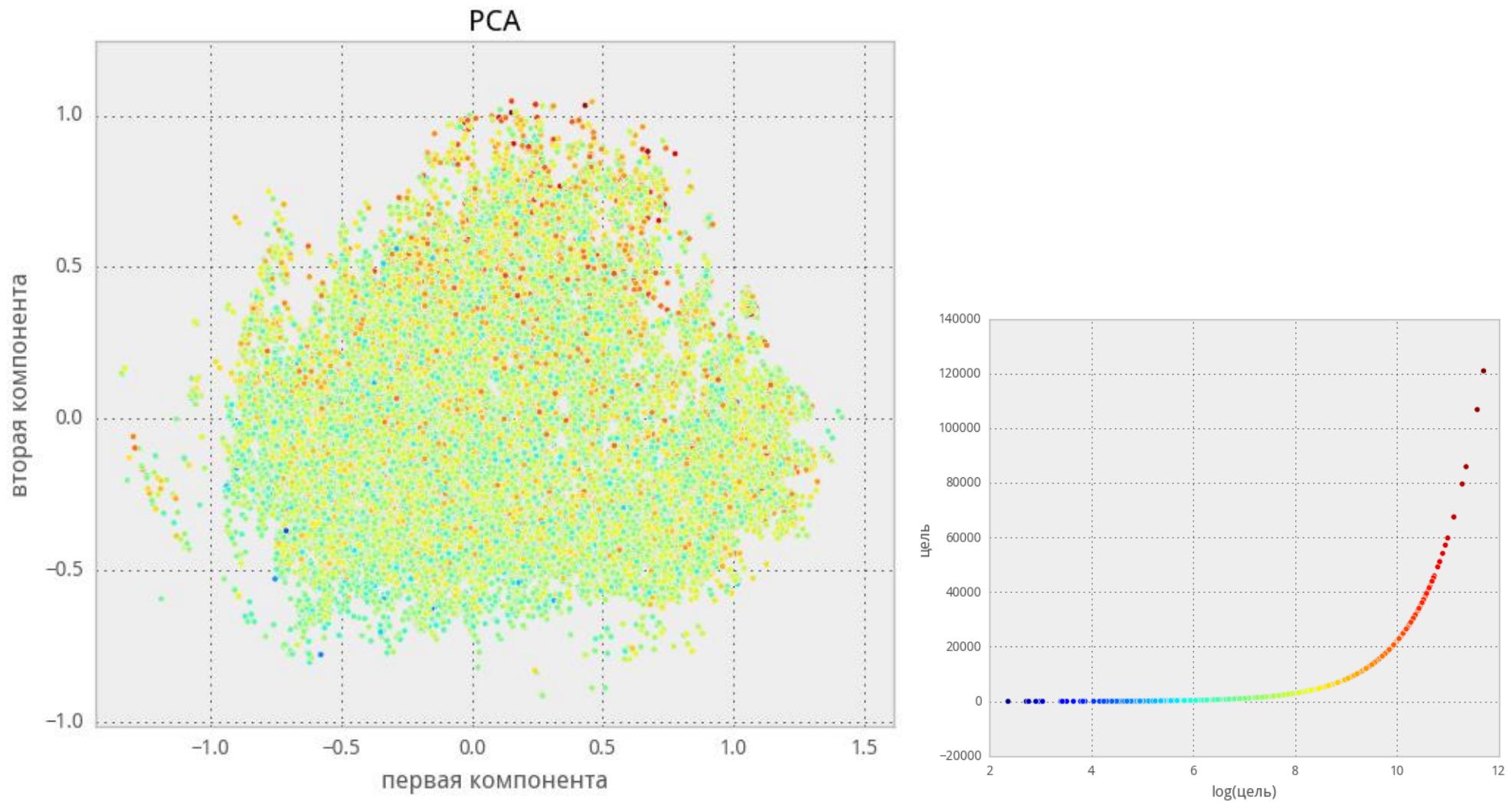
есть интересные сочетания



Смотрим на пары признаков



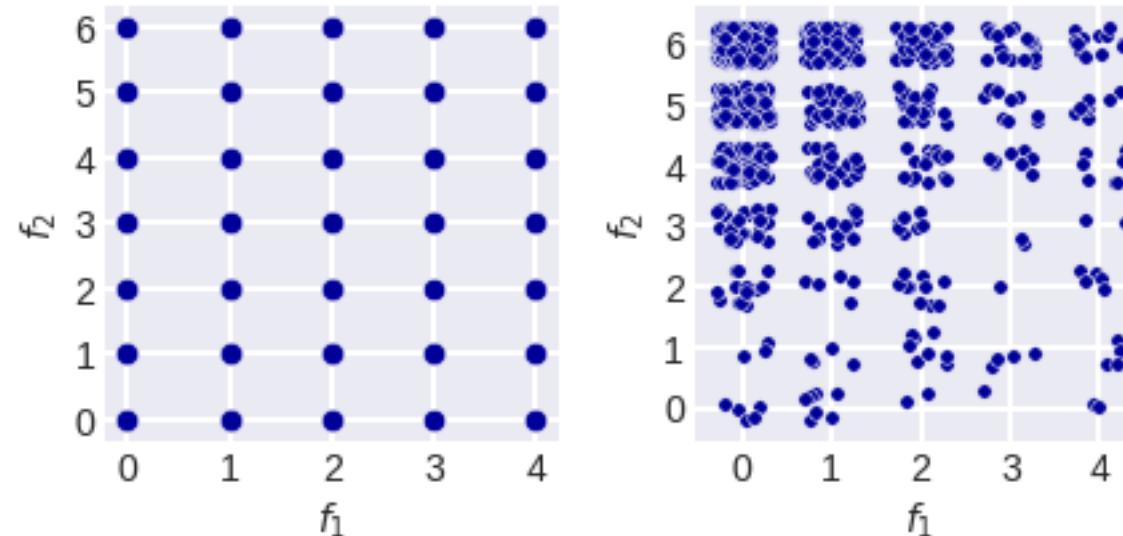
Визуализация сгенерированных признаков «AllState»



Что это за разложение Хорошее ли оно?

Диаграммы рассеивания дискретных признаков

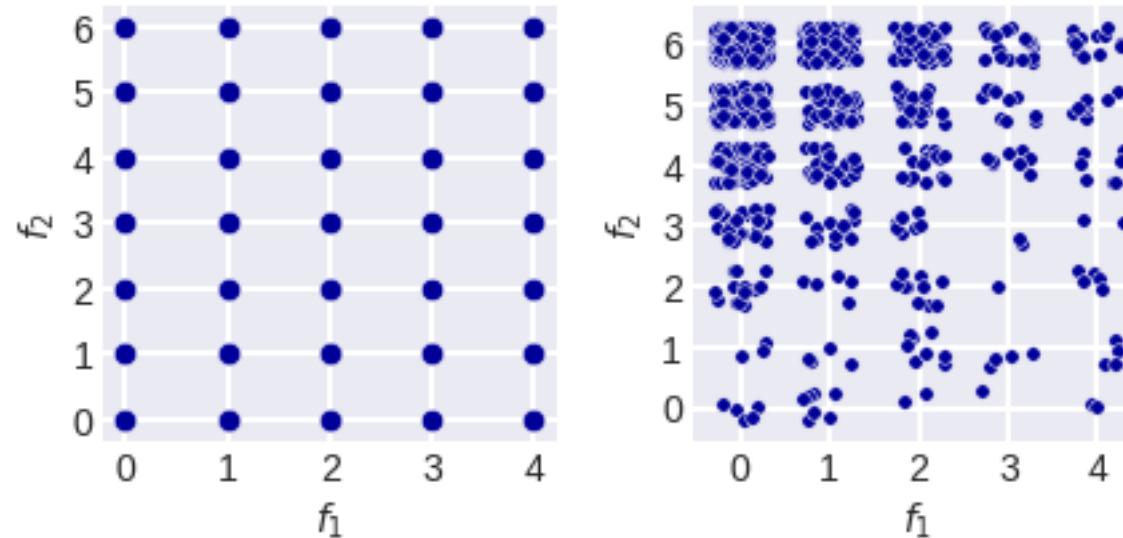
Зачем нужен Jitter



Что видно?

Диаграммы рассеивания дискретных признаков

Зачем нужен Jitter



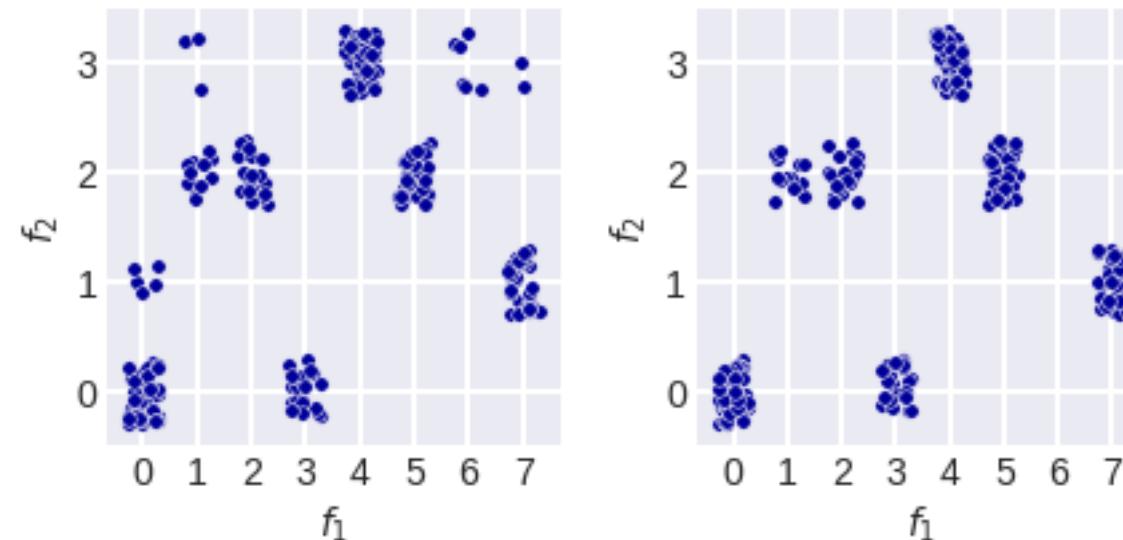
Что видно?

**«Треугольная зависимость»
(т.е. взаимная нумерация имеет смысл)**

Сводная таблица

	0	1	2	3	4	5	6
0	5	3	13	24	59	152	405
1	7	4	5	14	25	56	154
2	2	8	10	8	16	21	60
3	1	4	1	2	9	10	21
4	2	4	5	2	7	8	12

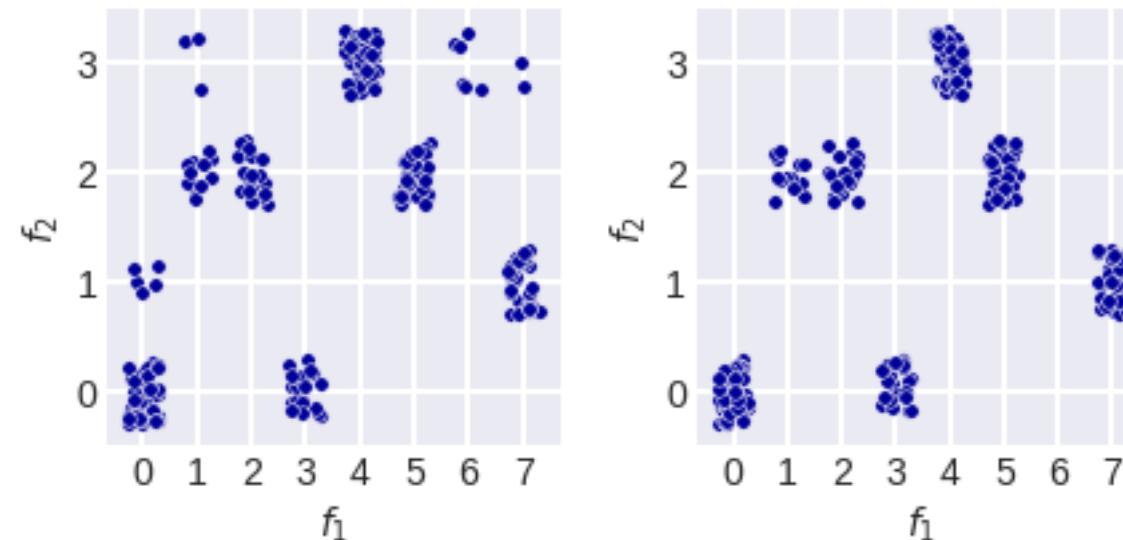
Диаграммы рассеивания дискретных признаков



Справа – после удаления маленьких кластеров!

Что здесь видно?

Диаграммы рассеивания дискретных признаков



**Один признак – уточнение другого!
Как это использовать?**

«Liberty»

Из задачи «Liberty»

Верхняя треугольная зависимость

```
table(train$T2_V6,train$T2_V14)
```

	1	2	3	4	5	6	7
1	9840	1463	831	376	106	28	17
2	485	21233	3957	4137	1440	396	128
3	79	141	2570	794	431	106	41
4	30	66	22	1180	204	175	75
5	9	15	7	3	212	58	60
6	0	6	0	1	2	96	53
7	0	4	1	4	2	0	115

Обоснование необходимости использования пар признаков

```
table(train$T2_V11,train$T2_V13)
```

	A	B	C	D	E
N	10160	323	803	513	2260
Y	100	191	6704	4571	25374

```
tapply(train$Hazard,  
list(train$T2_V11, train$T2_V13),  
mean)
```

	A	B	C	D	E
N	3.876378	5.099071	4.574097	5.518519	3.946460
Y	3.810000	4.319372	4.231653	4.175016	3.942815

Из задачи «RedHat»

```
people[:5]
```

	people_id	char_1	group_1	char_2	date	char_3	char_4	char_5	char_6	char_7	char_8	char_9	char_10
0	ppl_100	type 2	group 17304	type 2	2021-06-29	type 5	type 5	type 5	type 3	type 11	type 2	type 2	True
1	ppl_100002	type 2	group 8688	type 3	2021-01-06	type 28	type 9	type 5	type 3	type 11	type 2	type 4	False
2	ppl_100003	type 2	group 33592	type 3	2022-06-10	type 4	type 8	type 5	type 2	type 5	type 2	type 2	True
3	ppl_100004	type 2	group 22593	type 3	2022-07-20	type 40	type 25	type 9	type 4	type 16	type 2	type 2	True

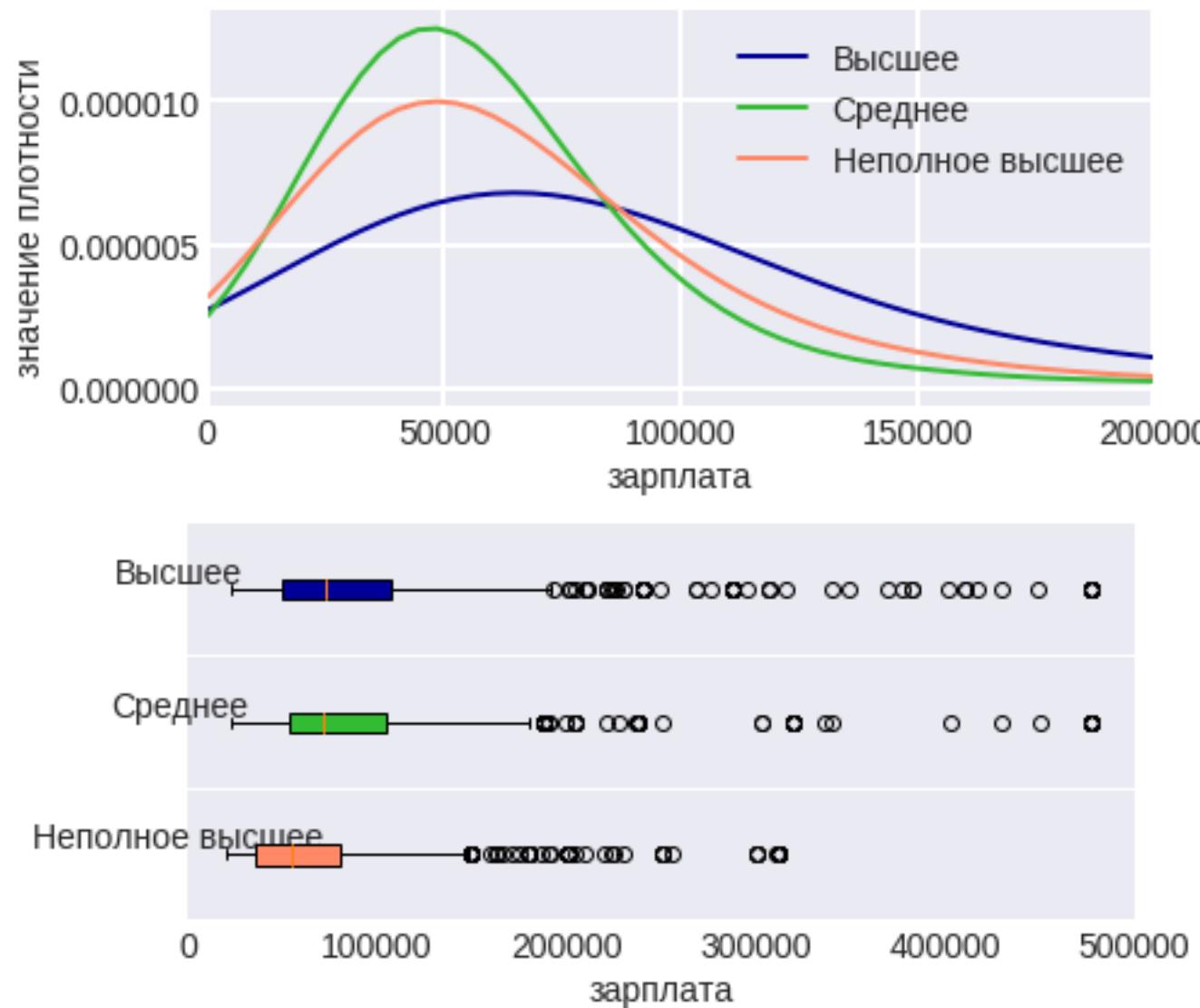
По таблице объект-признак сложно увидеть, что один категориальный признак – уточнение другого

```
pd.crosstab(people.char_1, people.char_2)
```

char_2	type 1	type 2	type 3
char_1			
type 1	15251	0	0
type 2	0	77314	96553

Как использовать это знание?

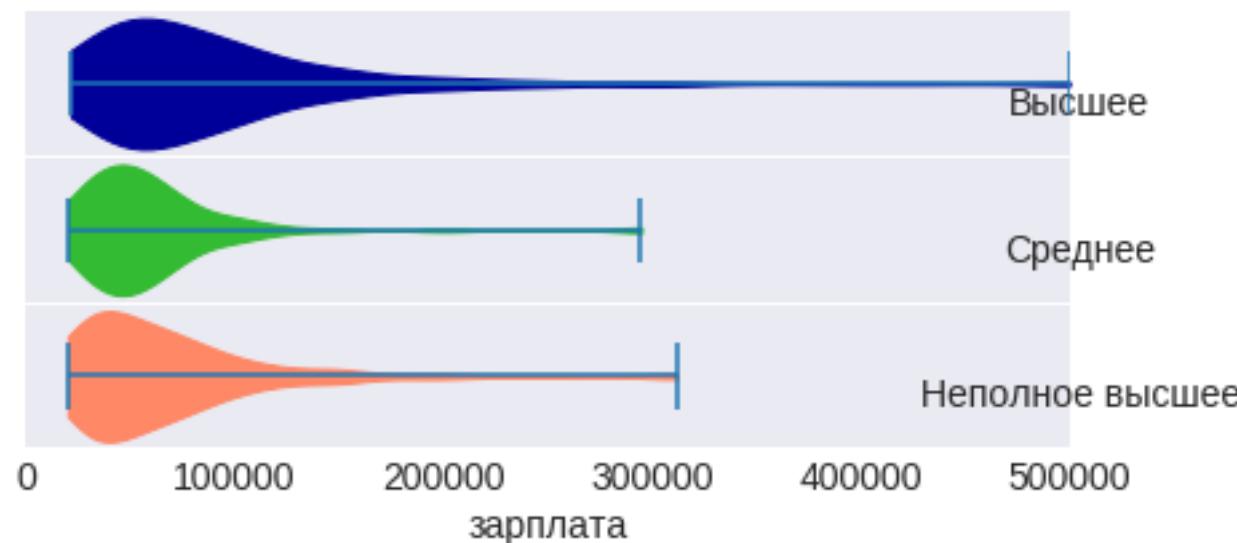
Пара «вещественный признак – категориальный»



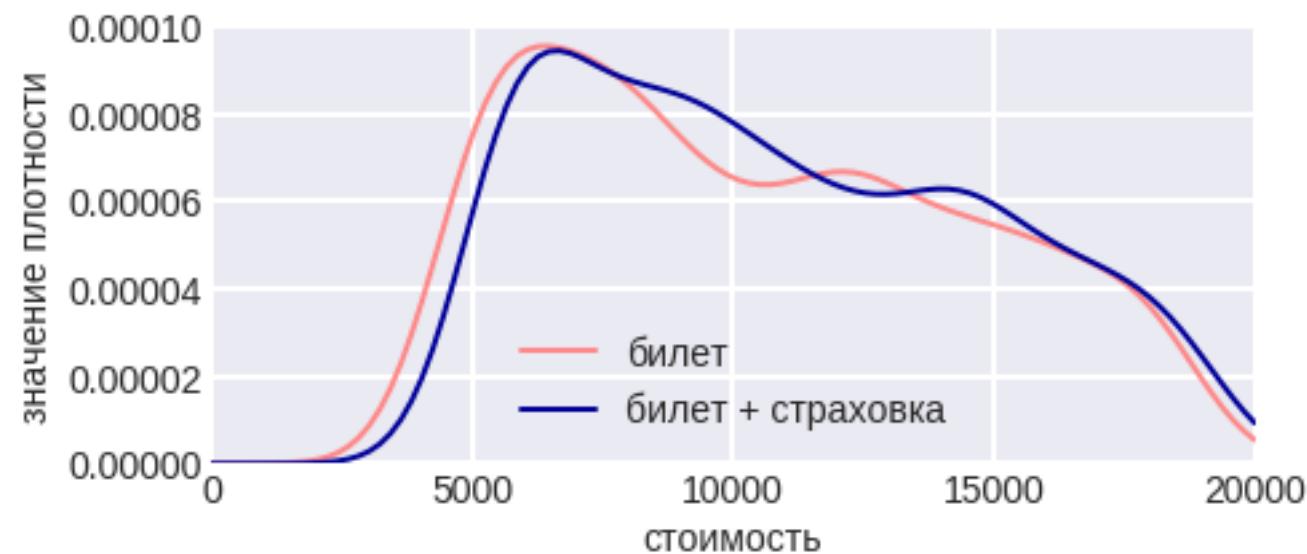
Ящик с усами (box-plot)



Всё это не очень наглядно...

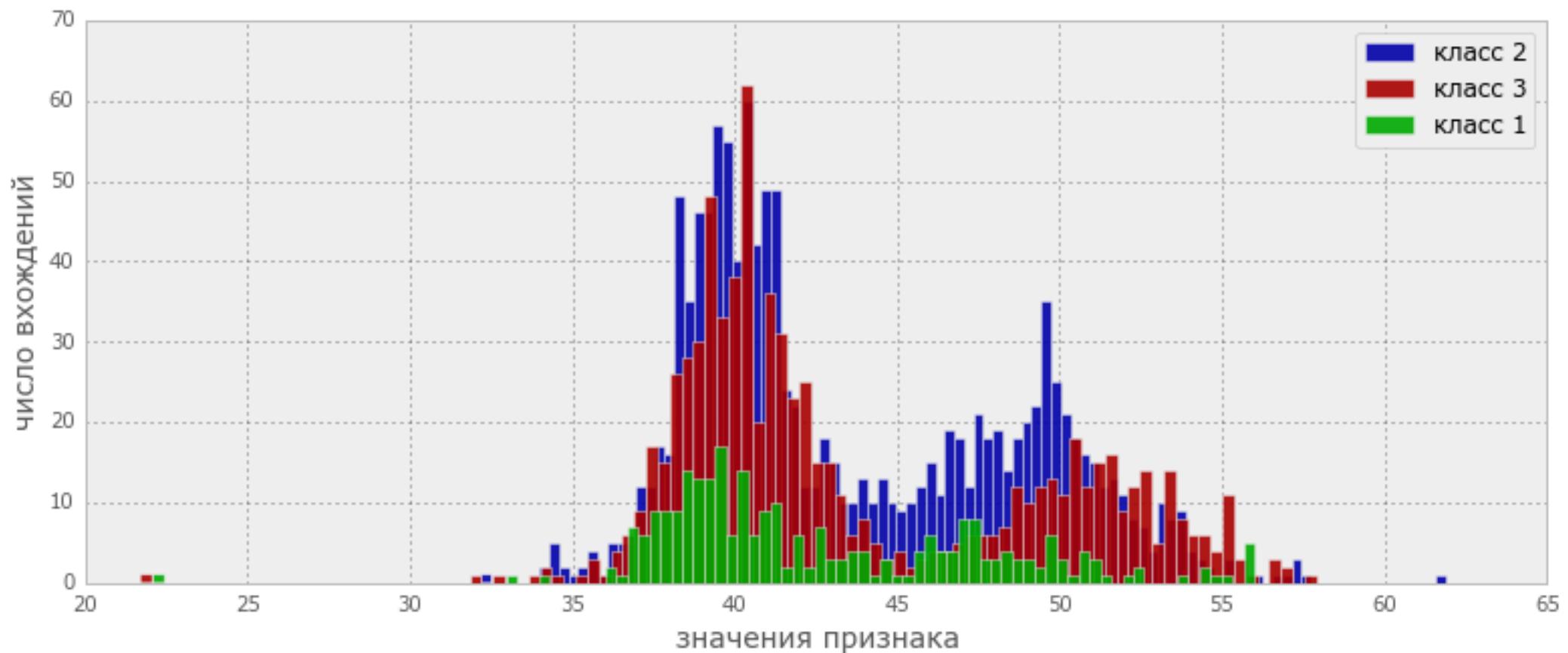


ЗАДАЧА «О-Т»

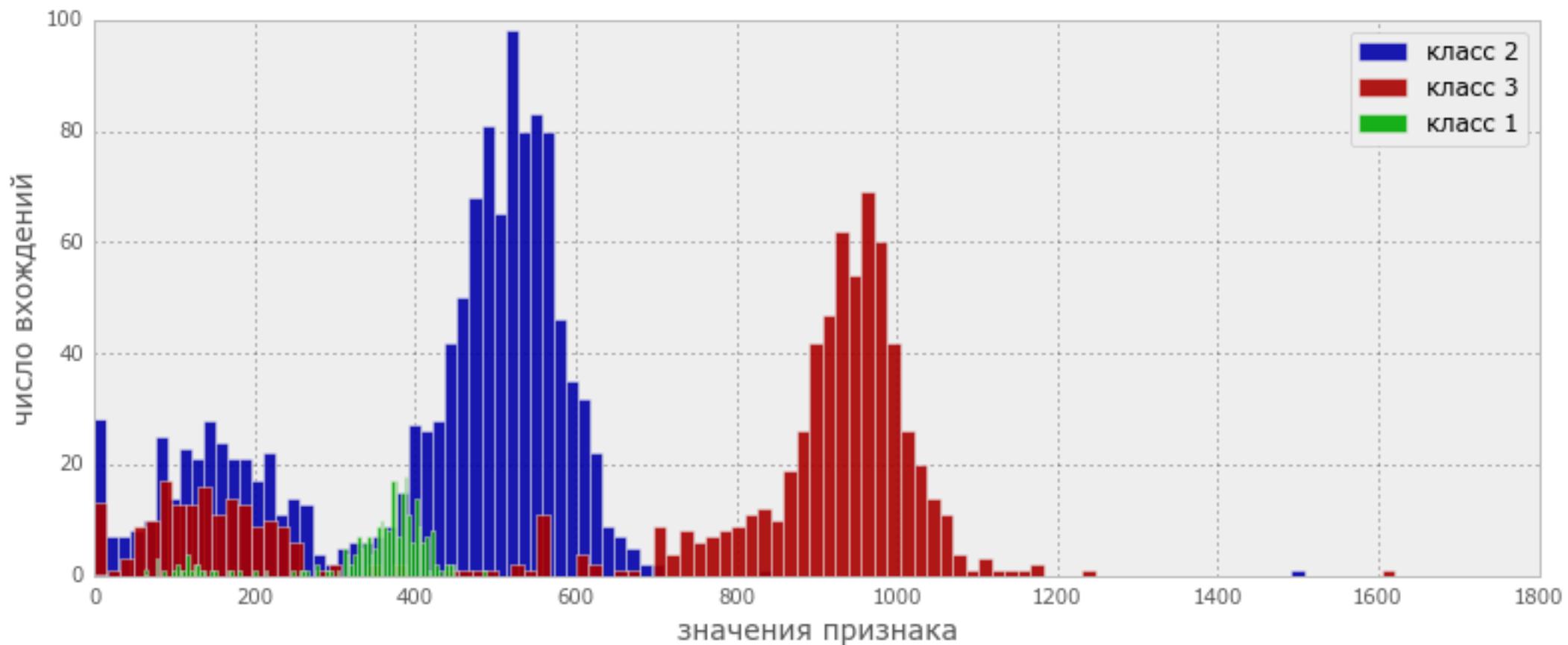


Всегда ставьте под сомнение свои выводы!

Как распределена цель на признаках

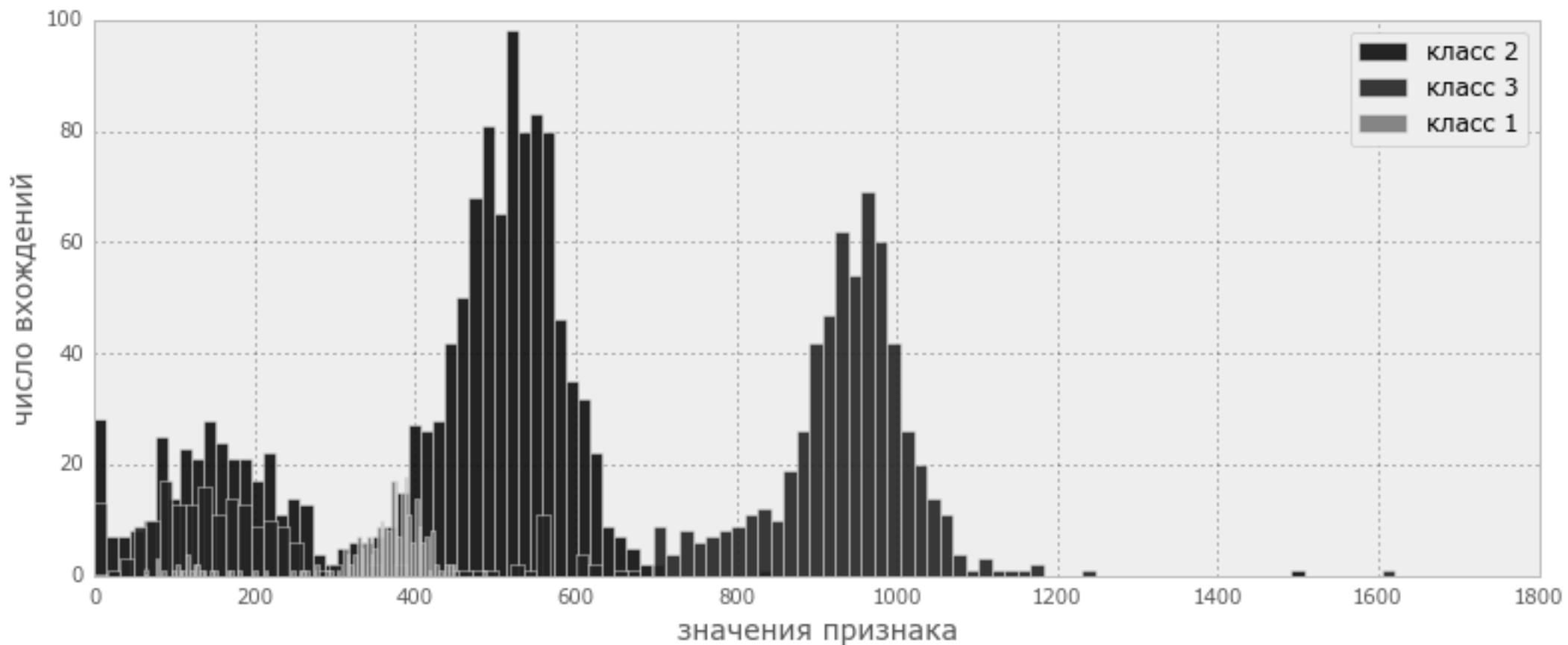


Как распределена цель на признаках



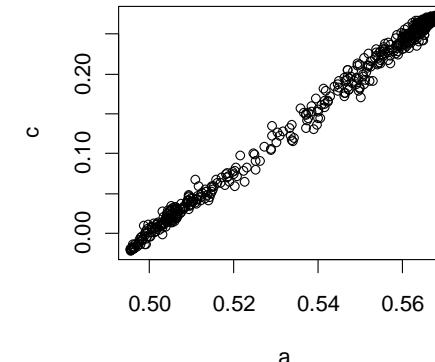
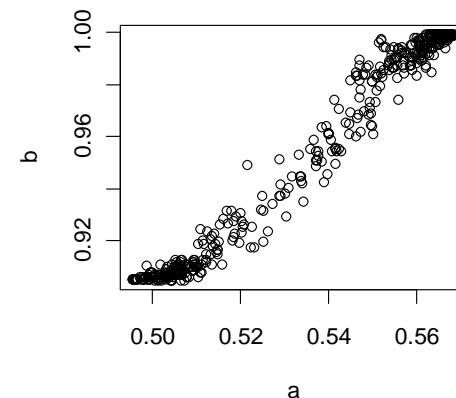
Чем плох рисунок?
Чем признак отличается от предыдущего?

Как распределена цель на признаках

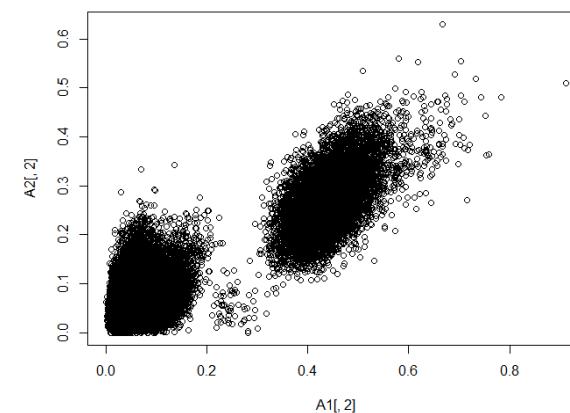
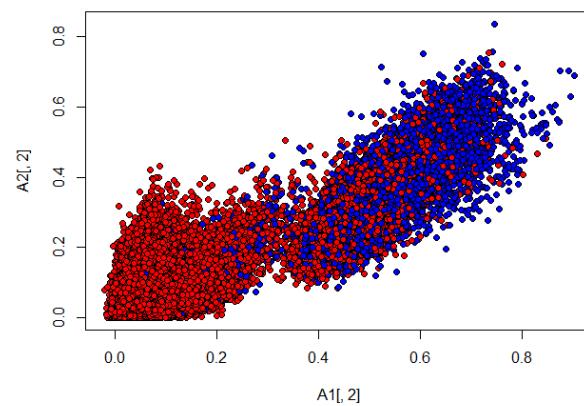


Вот чем...

Визуализация ответов двух алгоритмов



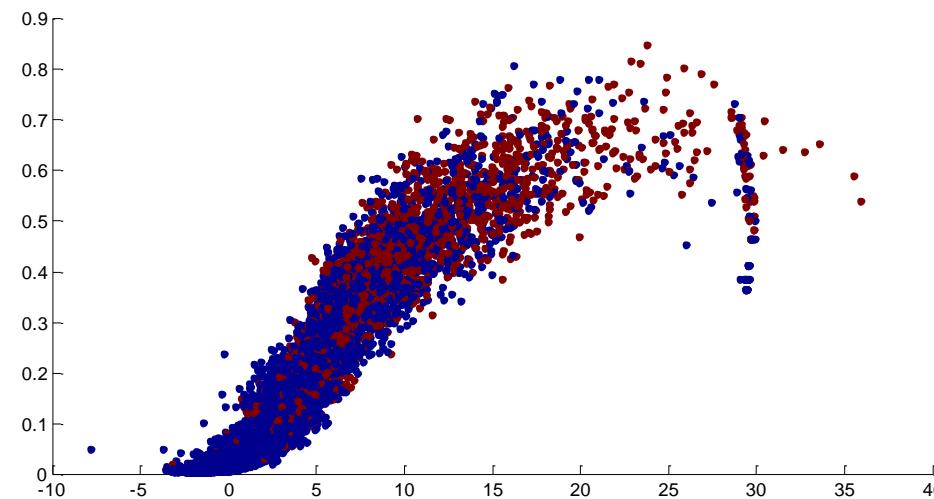
Как найти ошибку используя бенчмарк...



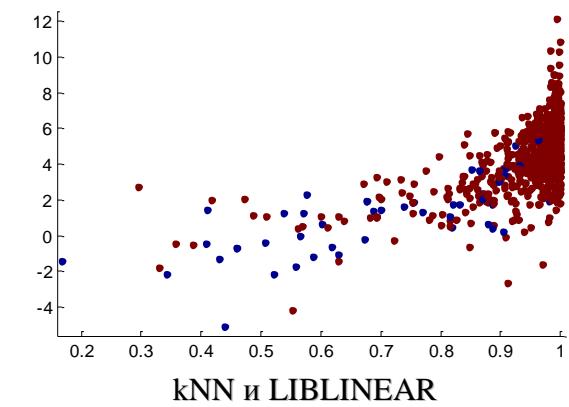
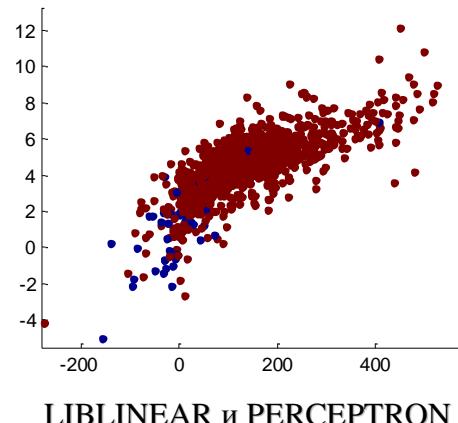
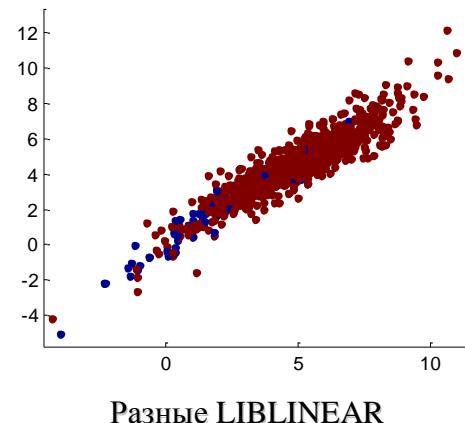
Совет: создавайте бенчмарк!

Ещё о визуализации «алгоритм-алгоритм»

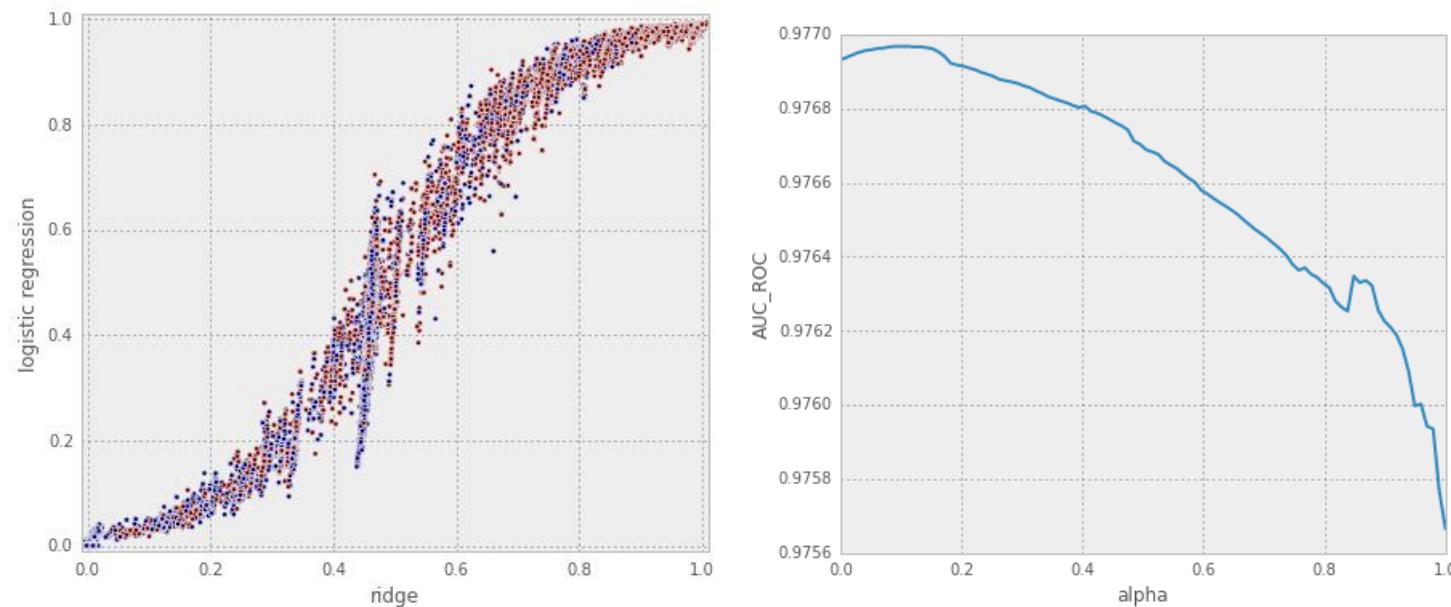
Задача скоринга



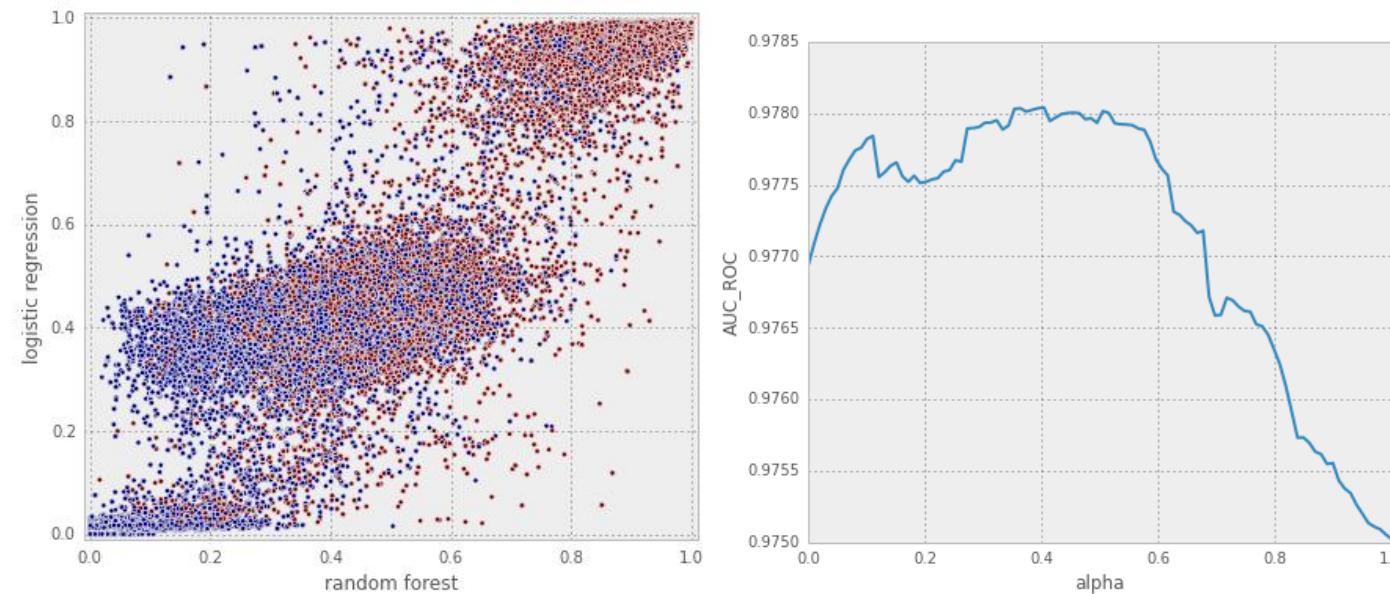
**Мой – горизонталь и RF – вертикаль
В задаче AMAZON**



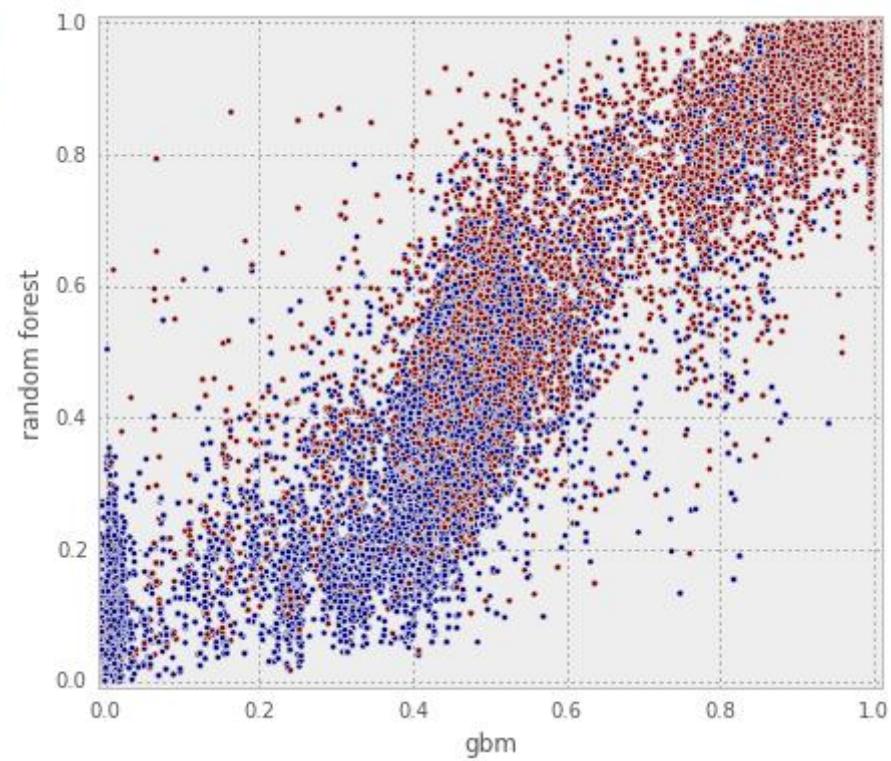
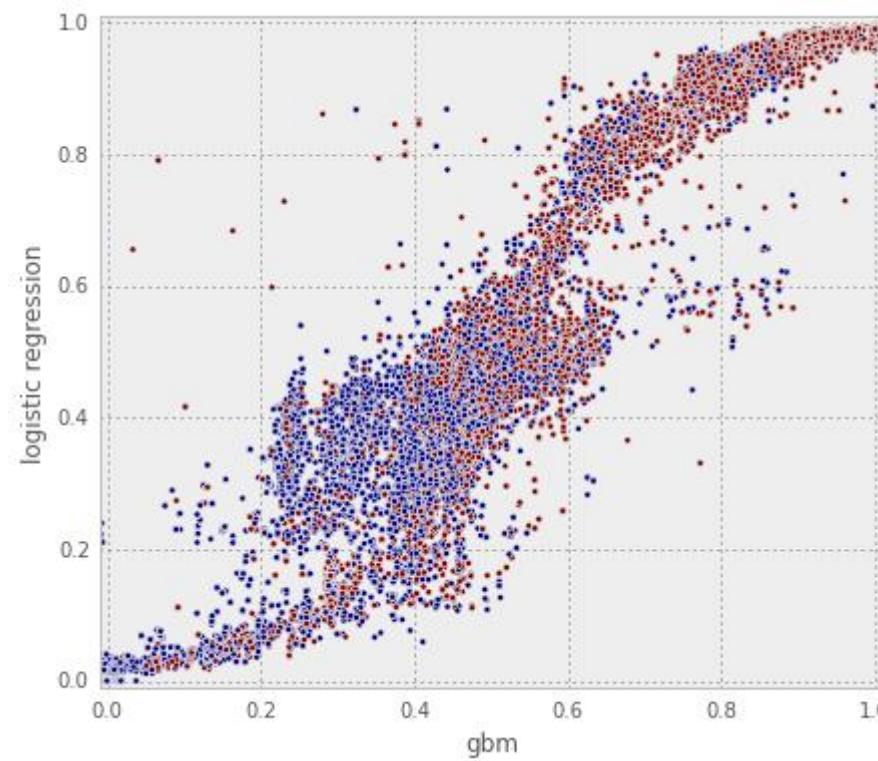
Ансамбль регрессия + логистическая регрессия



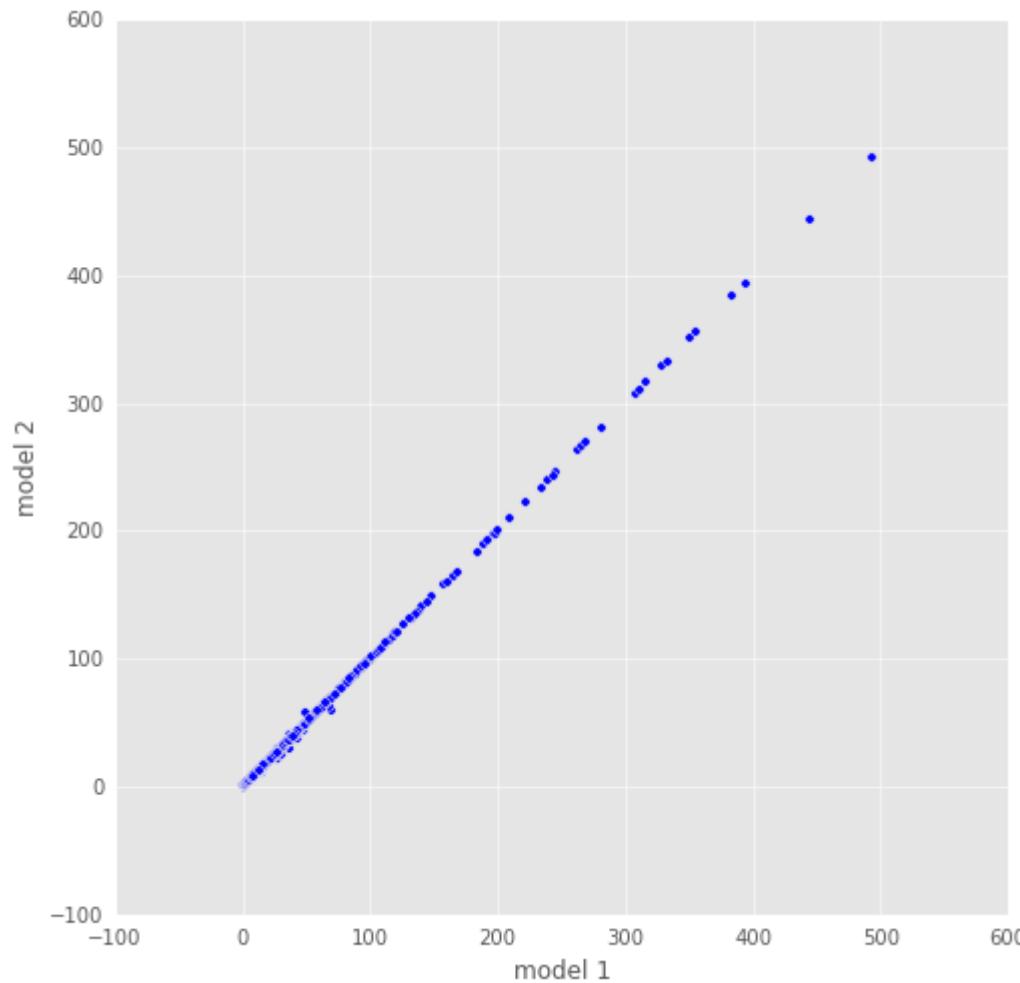
Ансамбль случайный лес + логистическая регрессия



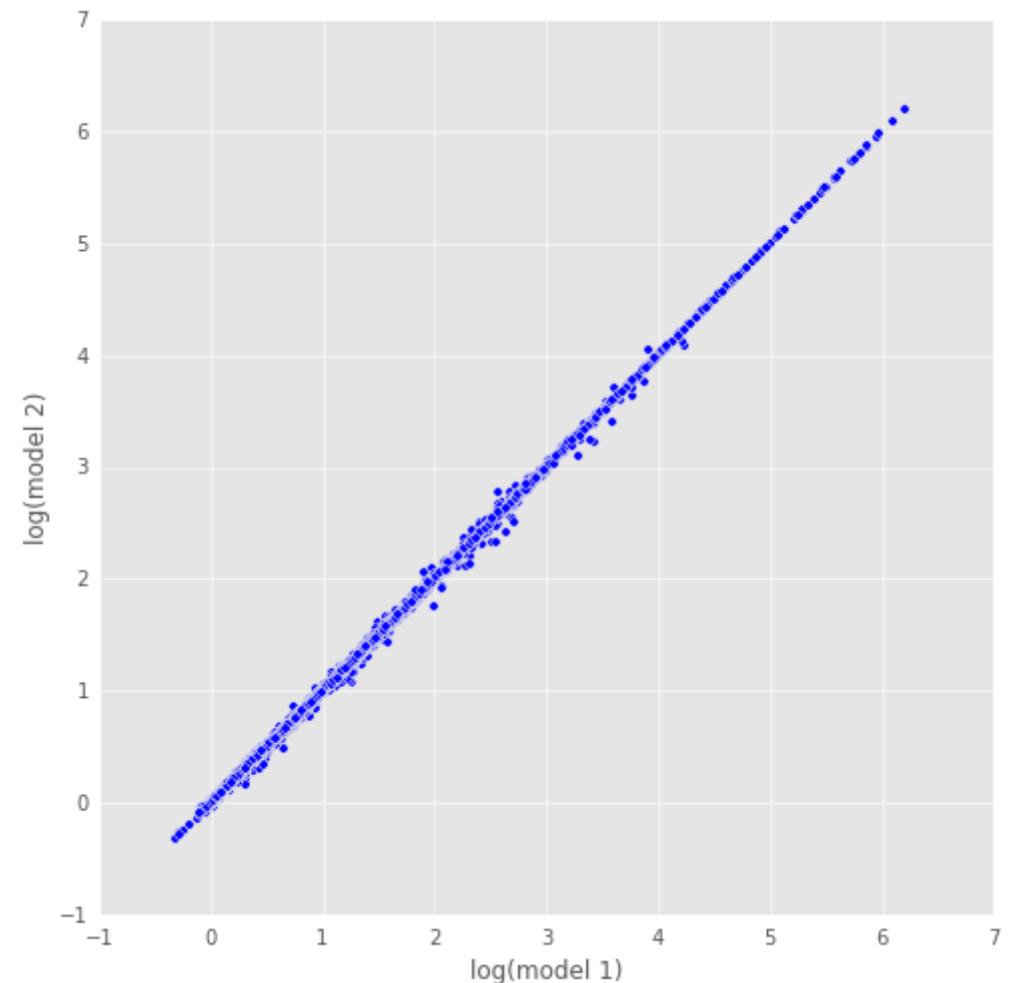
Ансамбли с gbm



Ещё о визуализации «алгоритм-алгоритм»

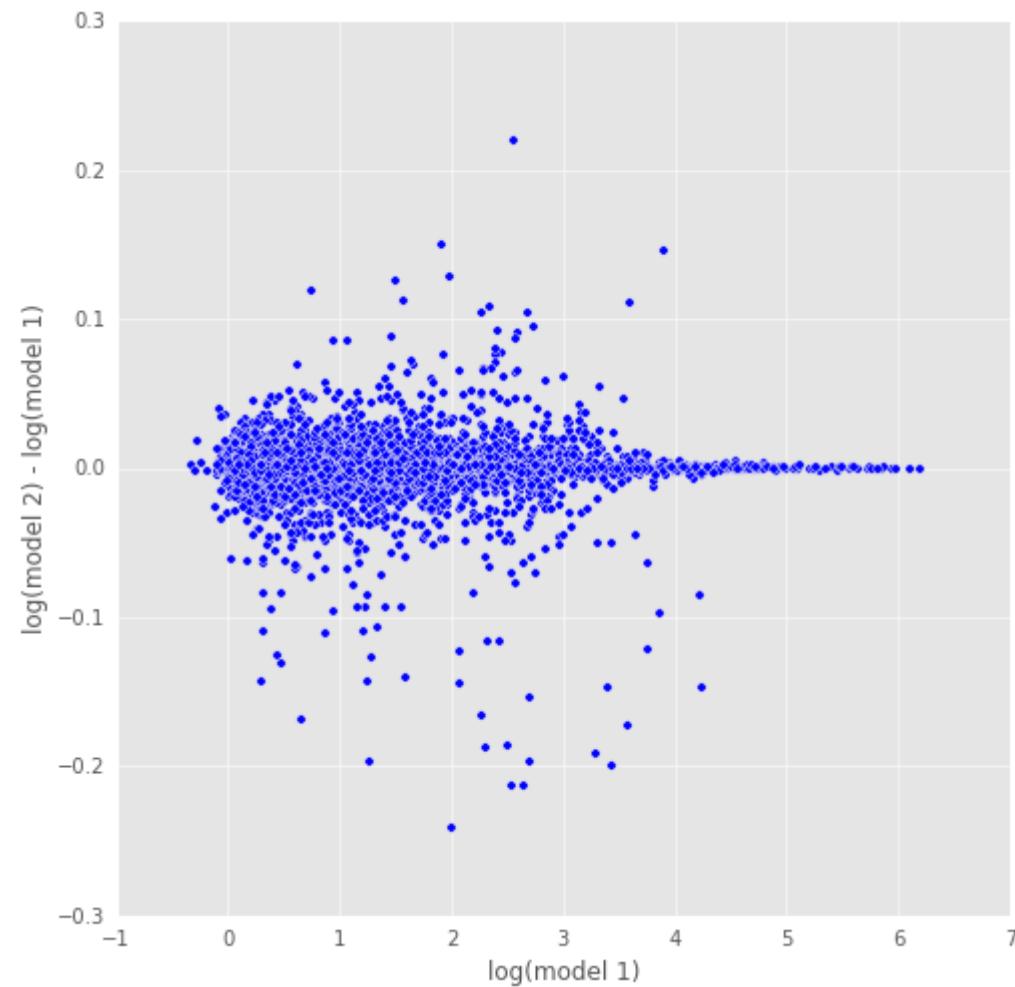


Две модели



Опять логарифмирование шкал!

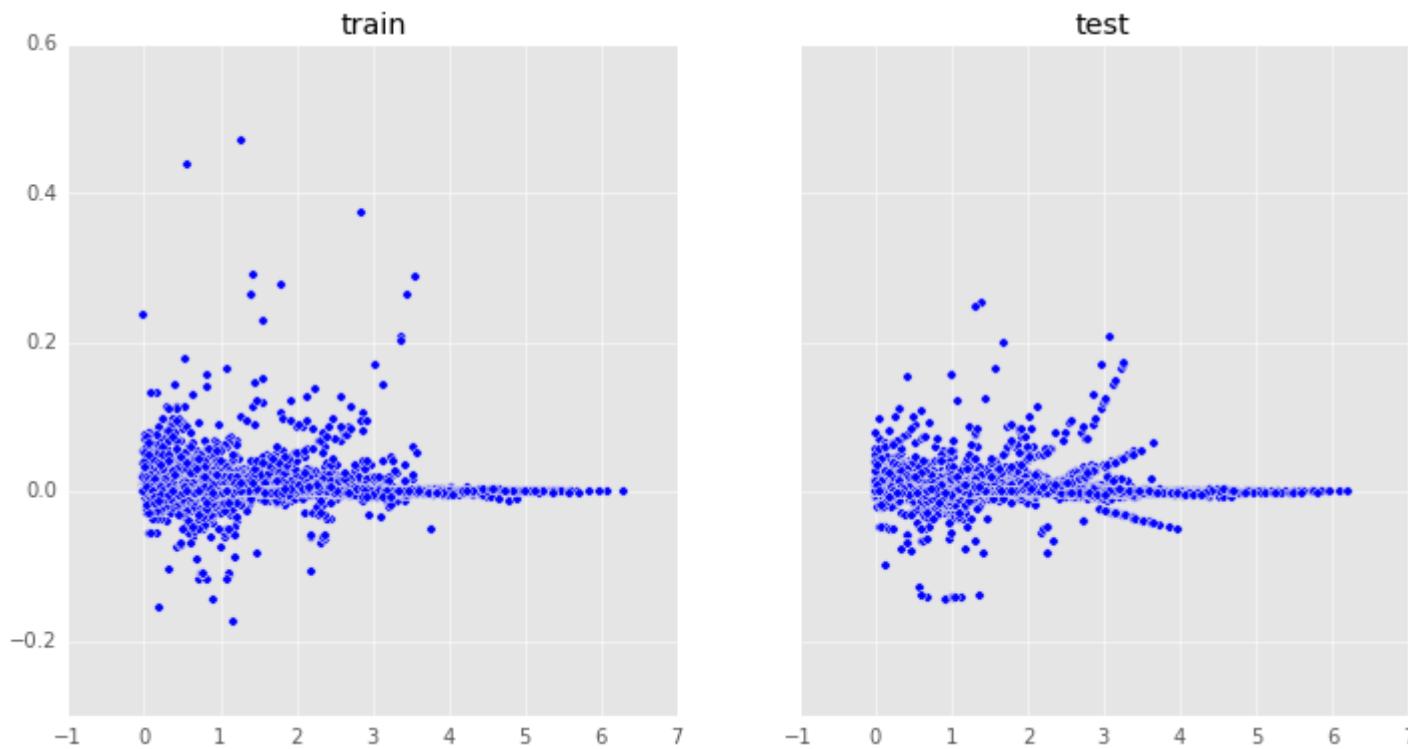
Ещё о визуализации «алгоритм-алгоритм»



Опять смотрим разницу ответов

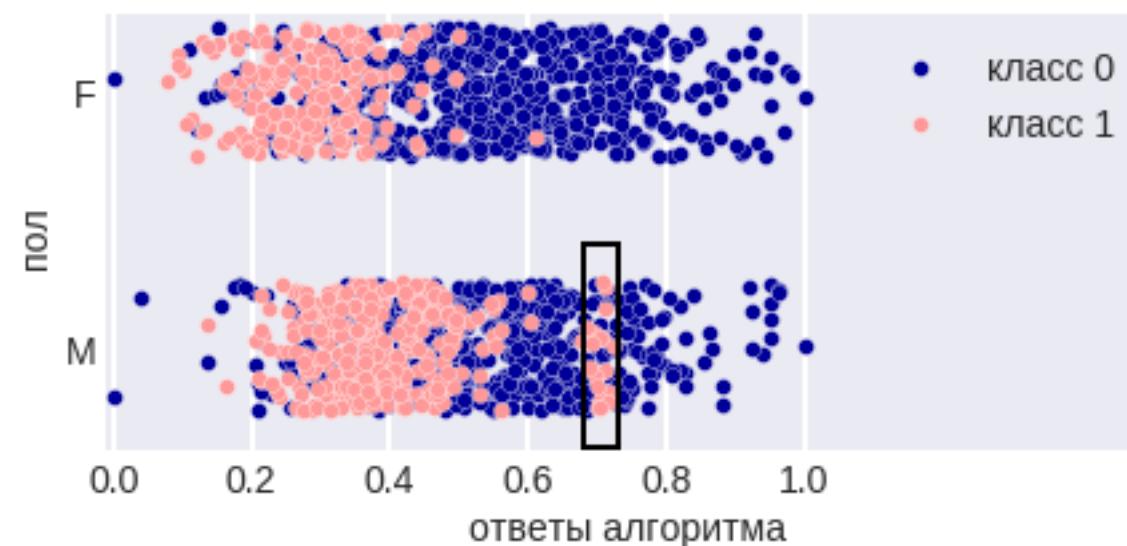
Наблюдение: при больших значениях модели работают идентично!

Ещё о визуализации «алгоритм-алгоритм»



На контроле подозрительные линии...
Что это может значить?
Что делать?

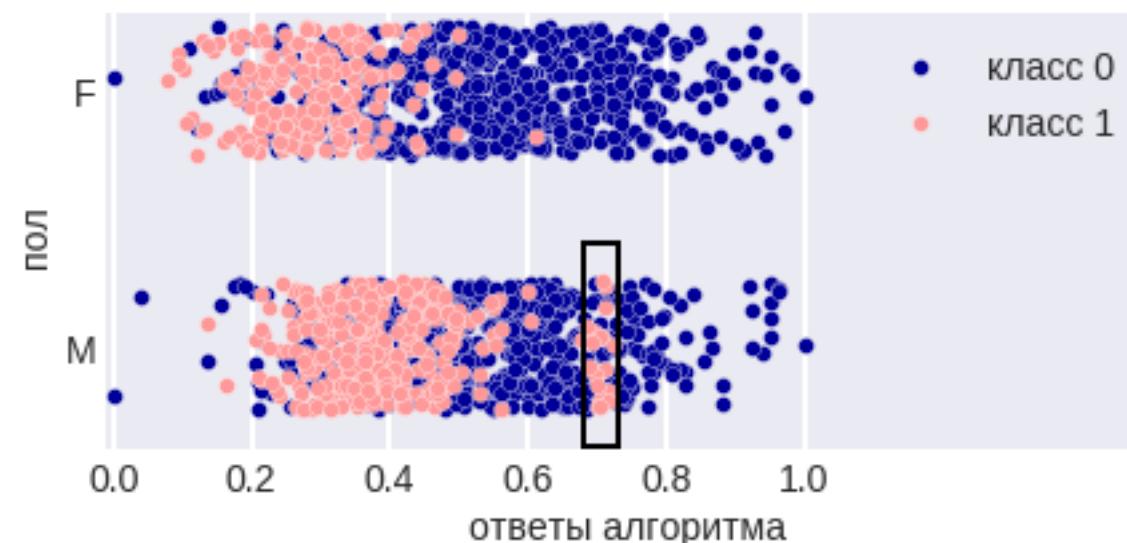
Ответы алгоритма – признак



Что видно?

Задача «~Analytics»

Ответы алгоритма – признак



Что видно:

- зона неверных ответов (почему?)
- порог зависит от значения признака «пол»

Но: распределения ответов зависит от контрольной выборки

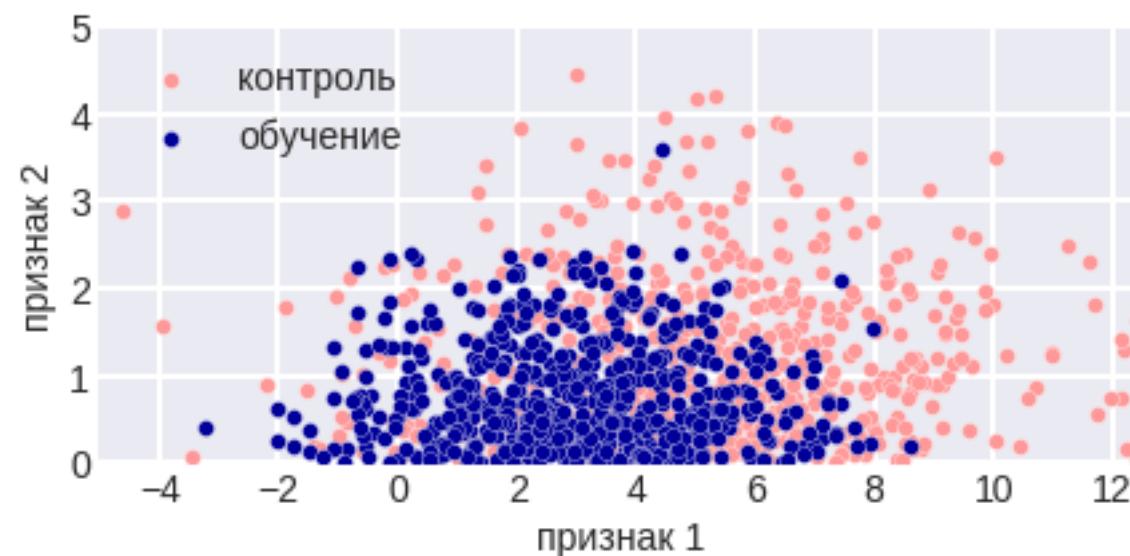
Что надо проверить найдя закономерность?

Что надо проверить найдя закономерность?

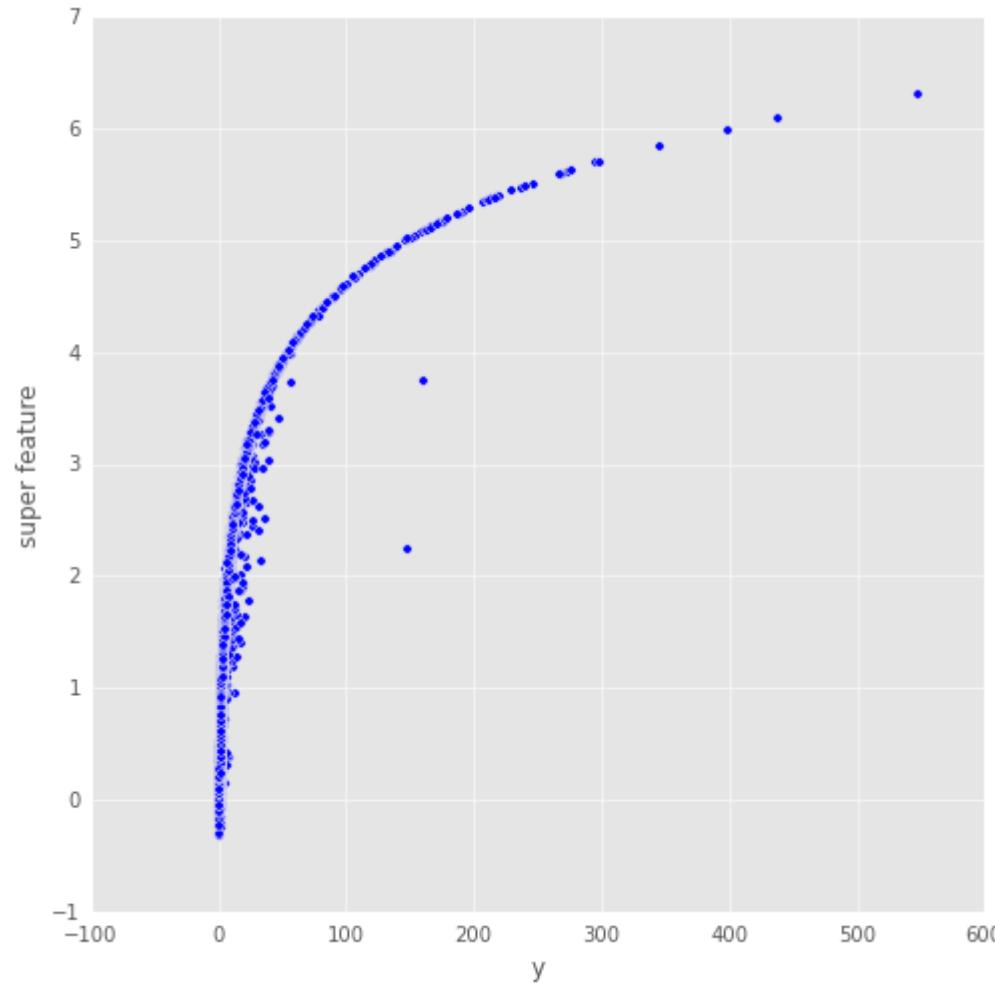
Что «контроль» ложится на обучение!

На практике нет гарантий одинаковости распределений гарантирует, даже если это гарантирует заказчик.

Примеры: рёбра в соцсети, заказы, разнесённые по времени (что-то приходится на праздники) и т.д.



Визуализация «алгоритм – признак» Что сделать, чтобы картинка стала понятнее?



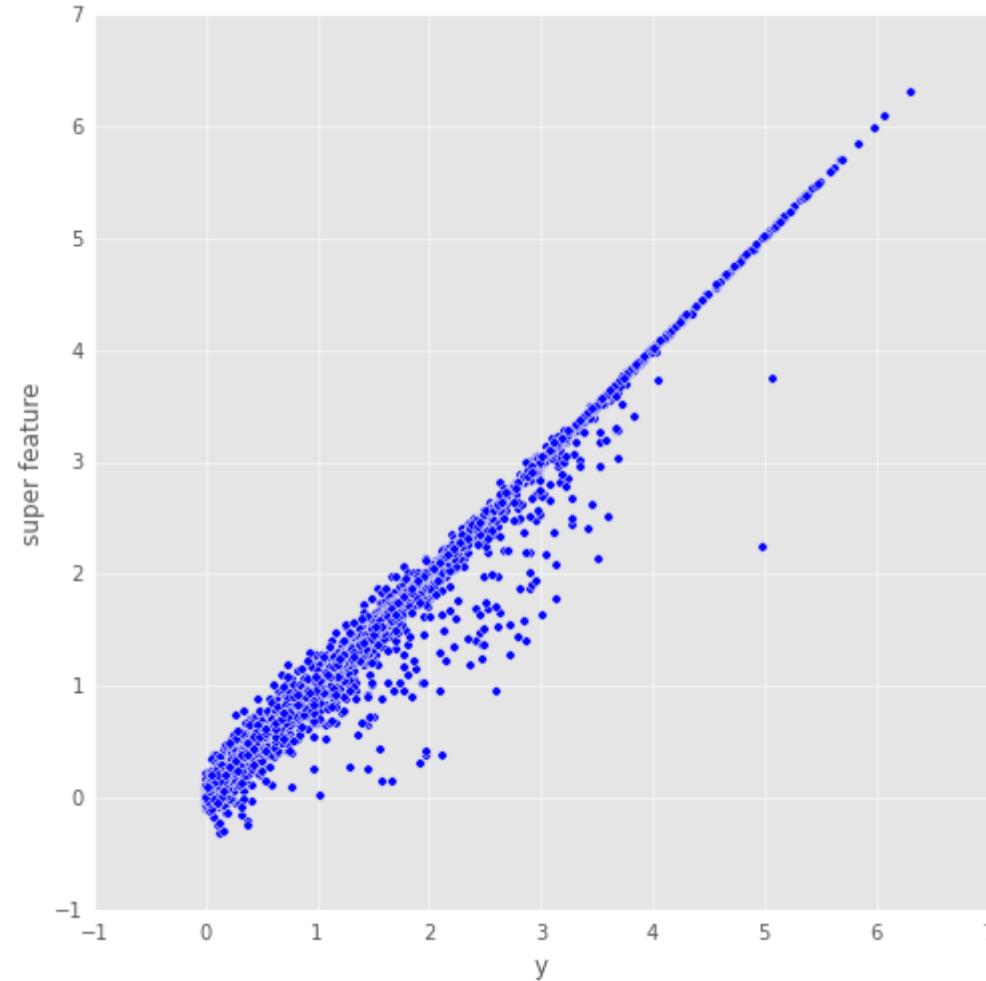
целевой признак и комбинация 2х признаков

Заметим, что эта комбинация строится как почти ответ...

```
plt.scatter((y2), np.log(train2.mnk.values) + train2.tmp.values)
```

Логарифмирование целевого признака

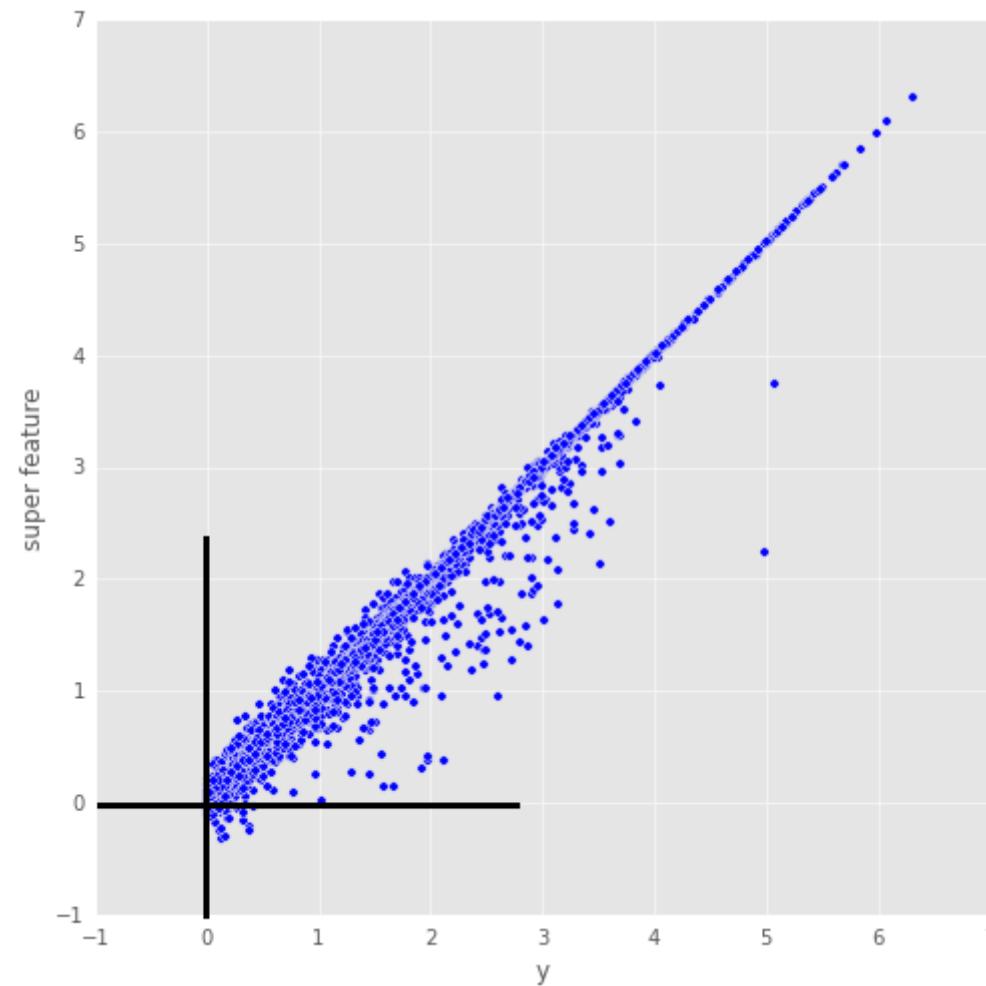
Что ещё **сделать**, чтобы картинка стала понятнее?



целевой признак и комбинация 2х признаков

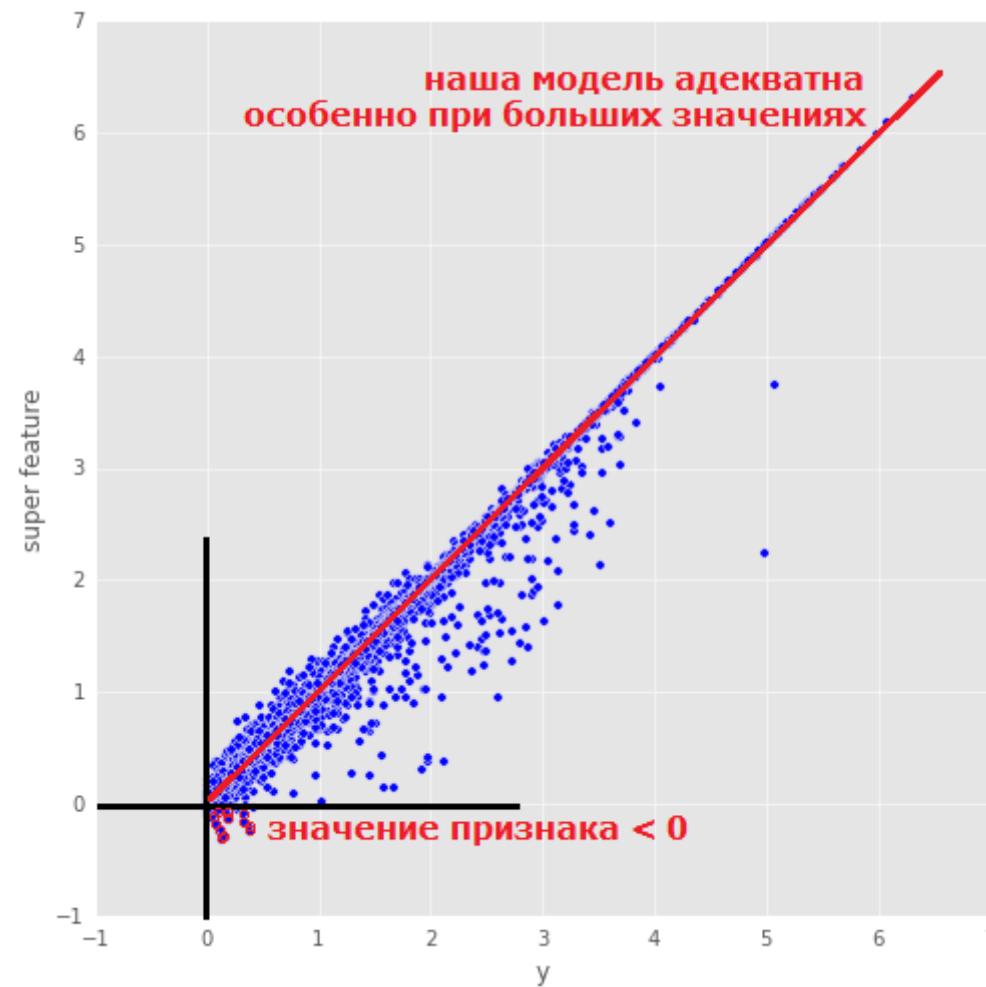
```
plt.scatter(np.log(y2), np.log(train2.mnk.values) + train2.tmp.values)
```

Логарифмирование целевого признака



Что видно на графике?

Логарифмирование целевого признака

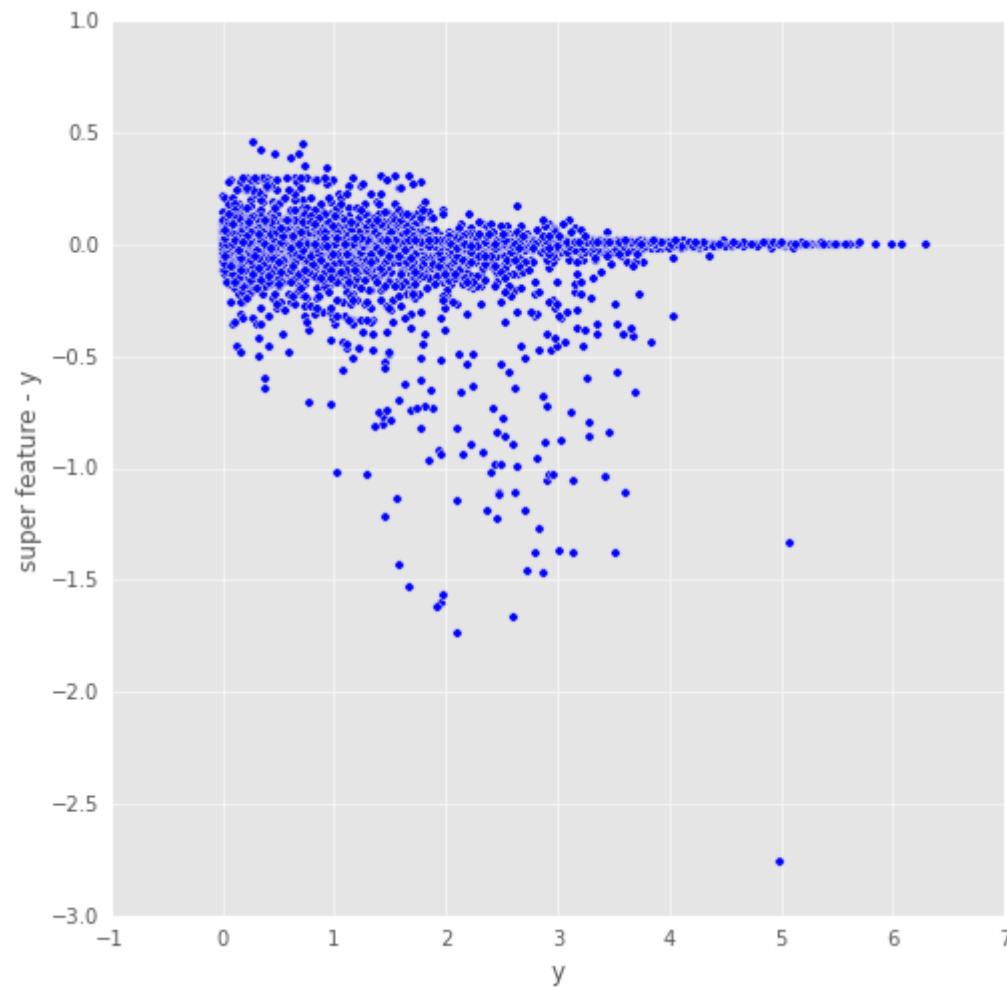


Правильный ответ всегда > 0

А наш супер-признак может принимать отрицательные значения!!!

Вывод: $\max(f, 0)$

Разница признака и целевого признака



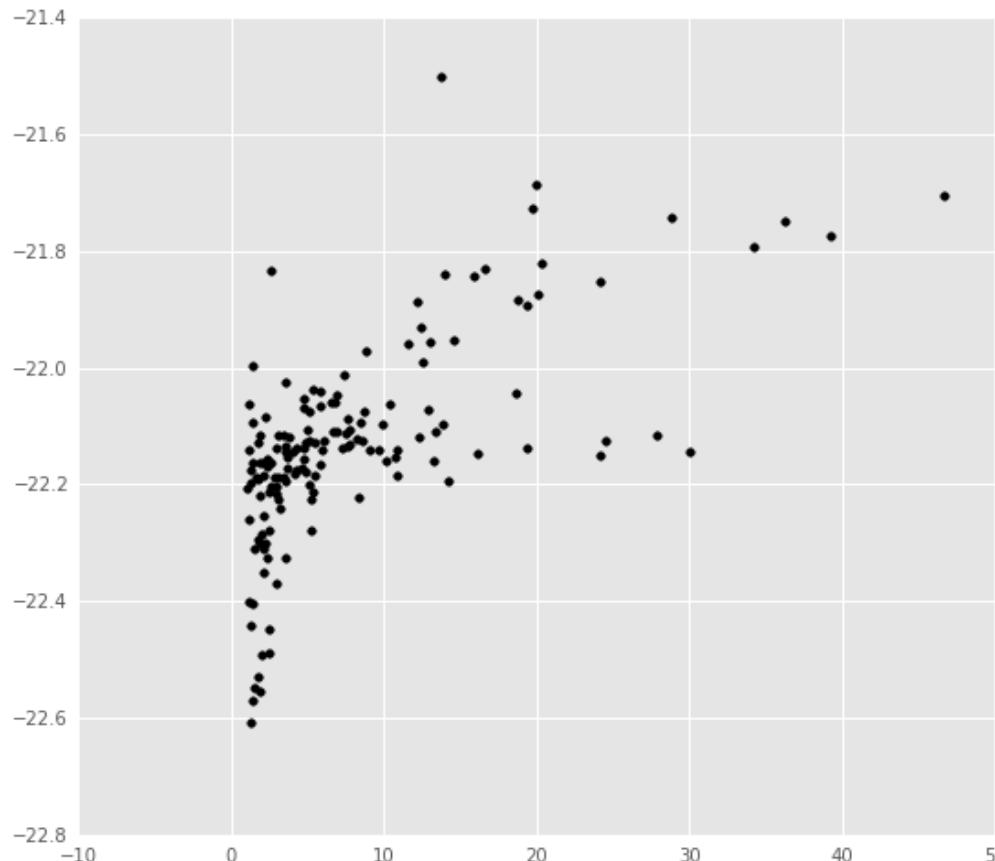
Если построили «почти ответ» – полезно посмотреть на ошибку

```
plt.scatter(np.log(y2), np.log(train2.mnk.values) + train2.tmp.values - np.log(y2))
```

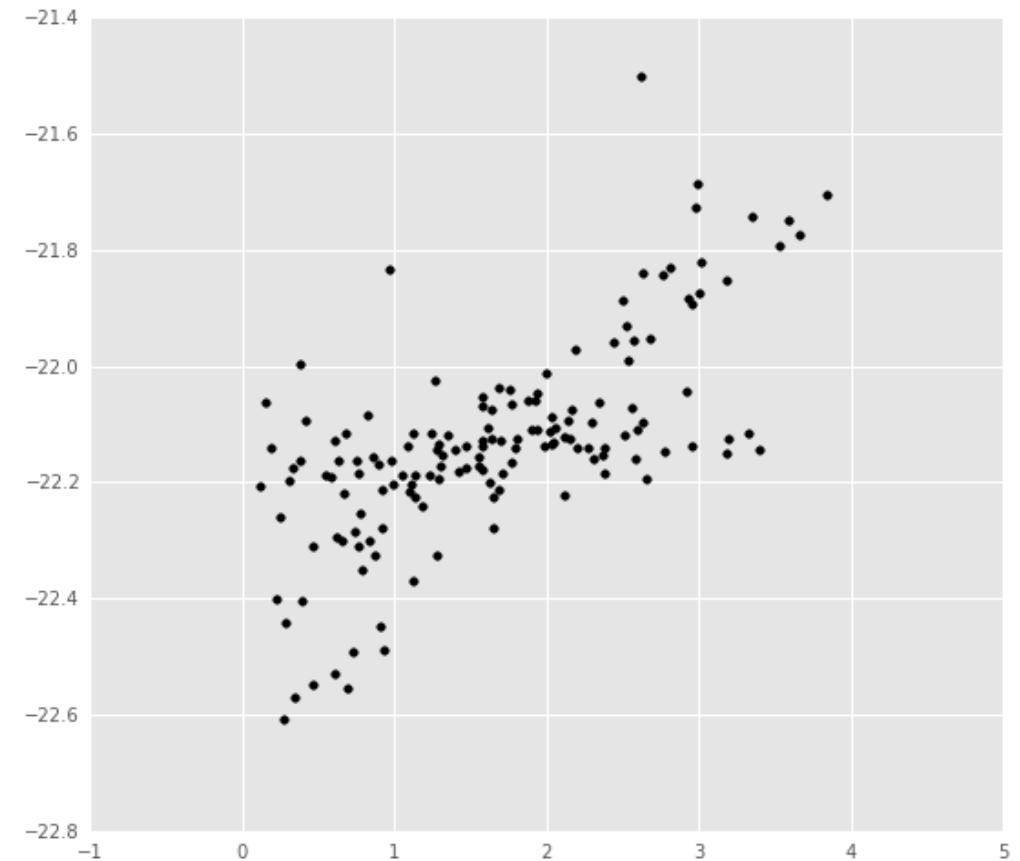
Residual plot

**диаграмма рассеивания «невязка» (?) – «прогнозируемая величина»
могут подсказать нужную трансформацию**

Необходимость логарифмирования можно не заметить на меленьких выборках

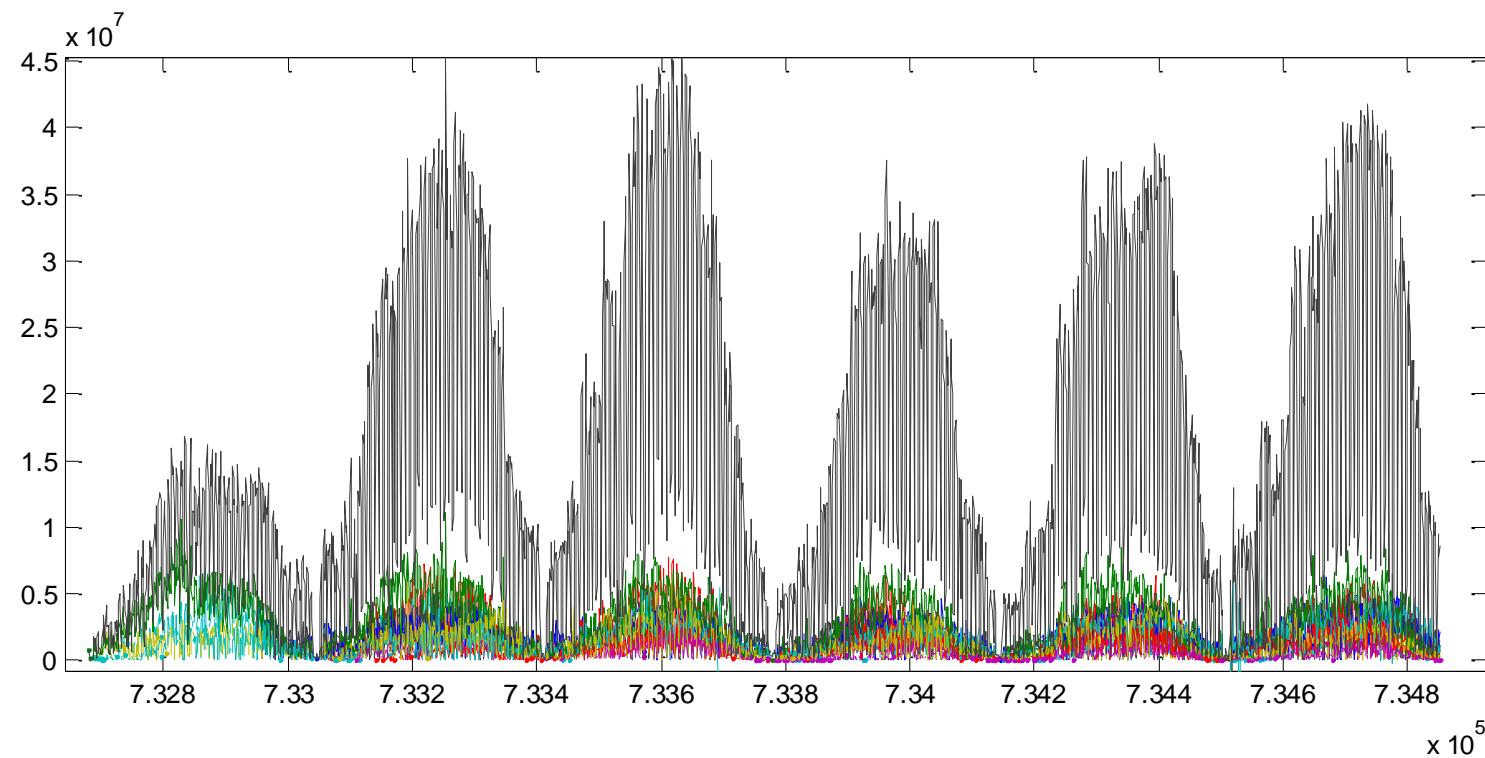


До логарифмирования



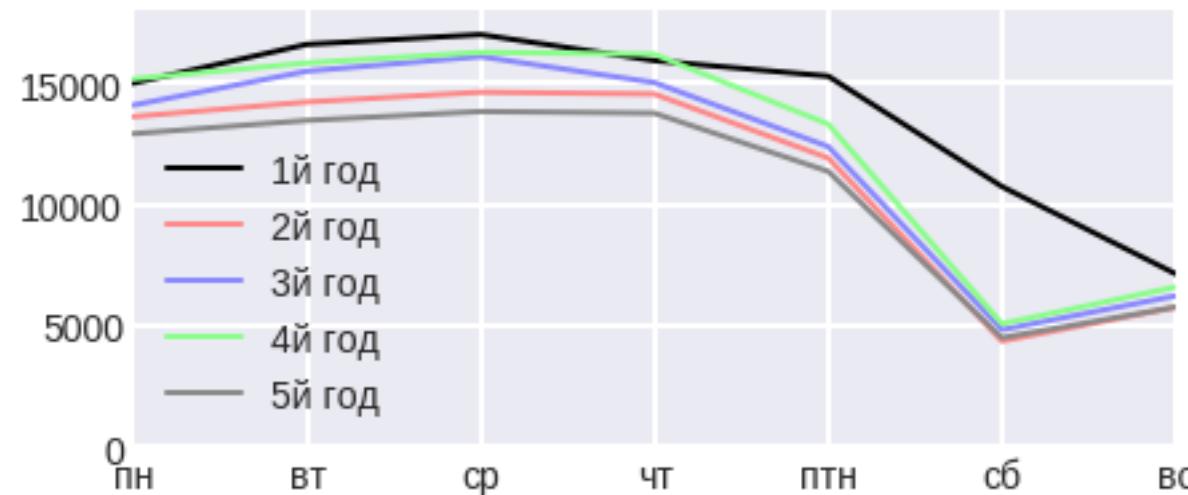
после

Агрегация (по дням недели) прогнозирование временного ряда (продажи)



Есть отрицательные значения – выбросы вниз (!?).

Агрегация (по дням недели)



Первый год нетипичен!

Остальные – очень похожи... осталось научиться прогнозировать «уровень недели».

Агрегация

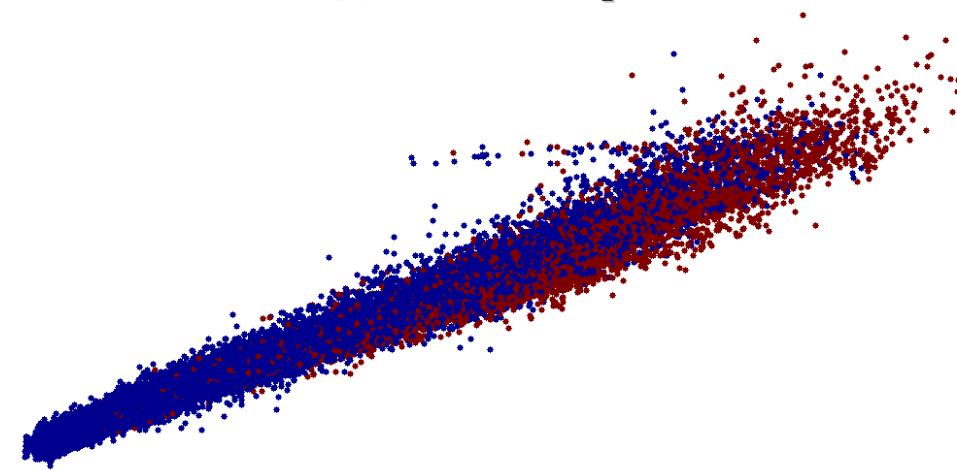
**Типичная ошибка:
что агрегировать**

- **все покупки (проблема оптовиков)**
- **средние покупки всех пользователей**

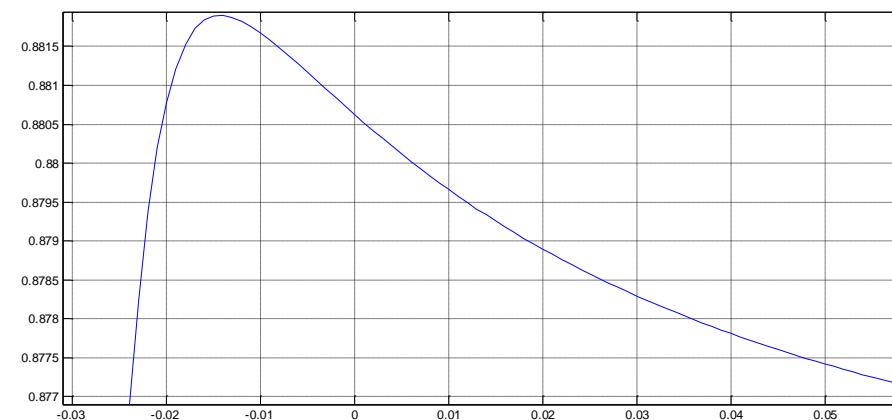
**Прошлый год – задача Сбербанка
«мужские» / «женские» товары**

Одномерная визуализация: качество алгоритма от параметра

Задача скоринга



Байес и (RF+GBM)



Коэффициент в линейной комбинации. Лучше вычитать!

Удивительно, но при визуализации:

- гладкость
- монотонность или унимодальность
- м.б. + явные выбросы

Если этого нет:

- ищем ошибку

3D-визуализации

Третий признак

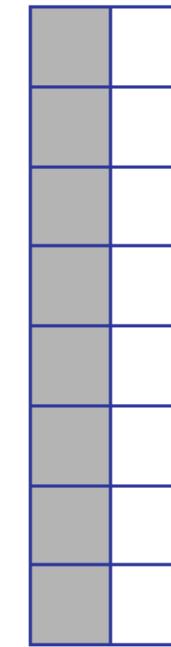
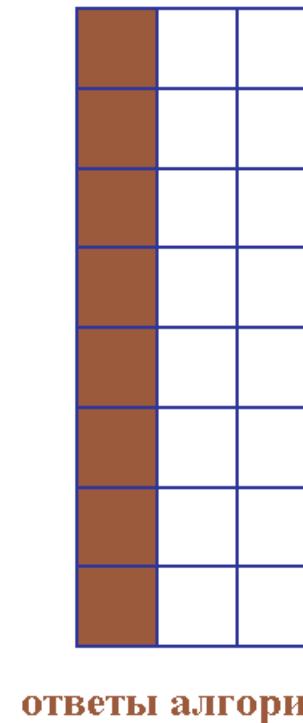
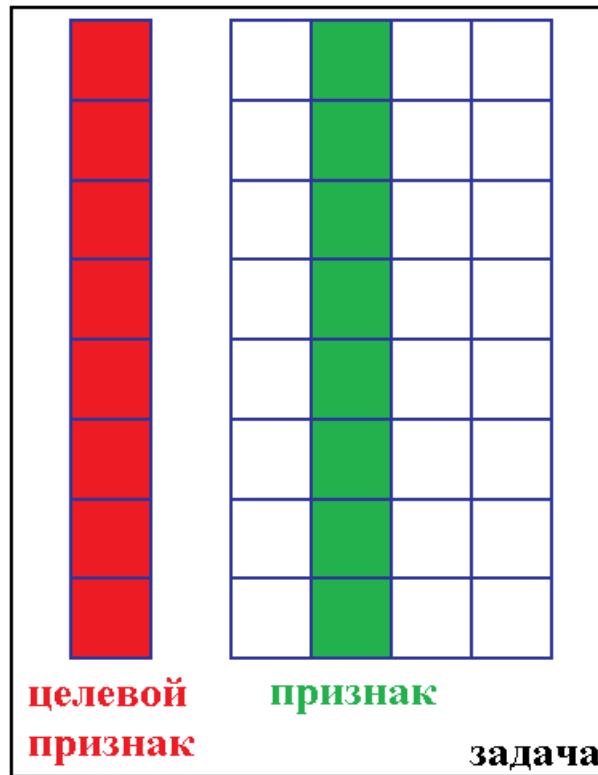
- цвет
- размер
- форма

Практически не делают!

Иногда, если объектов мало и можно интерактивно вращать

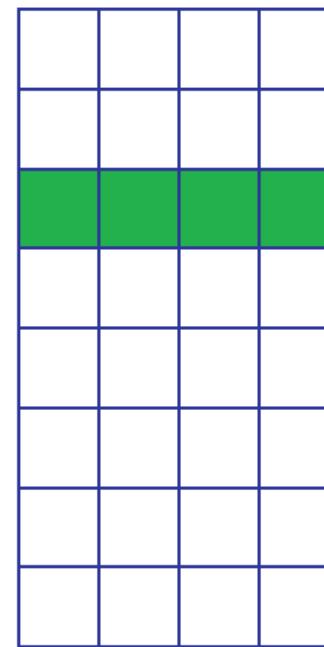
Что можно визуализировать:

«Всё вертикальное»



Что можно визуализировать:

«Всё горизонтальное» (реже)



объект

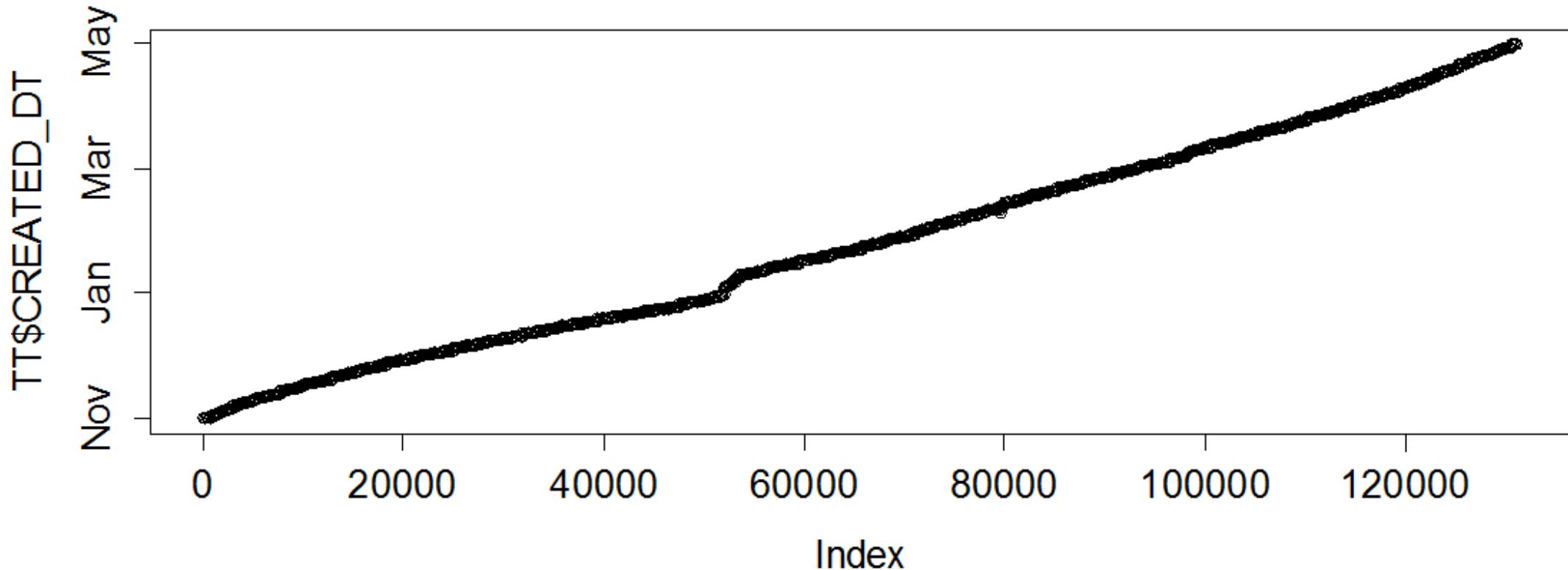


статистики признаков



dummy

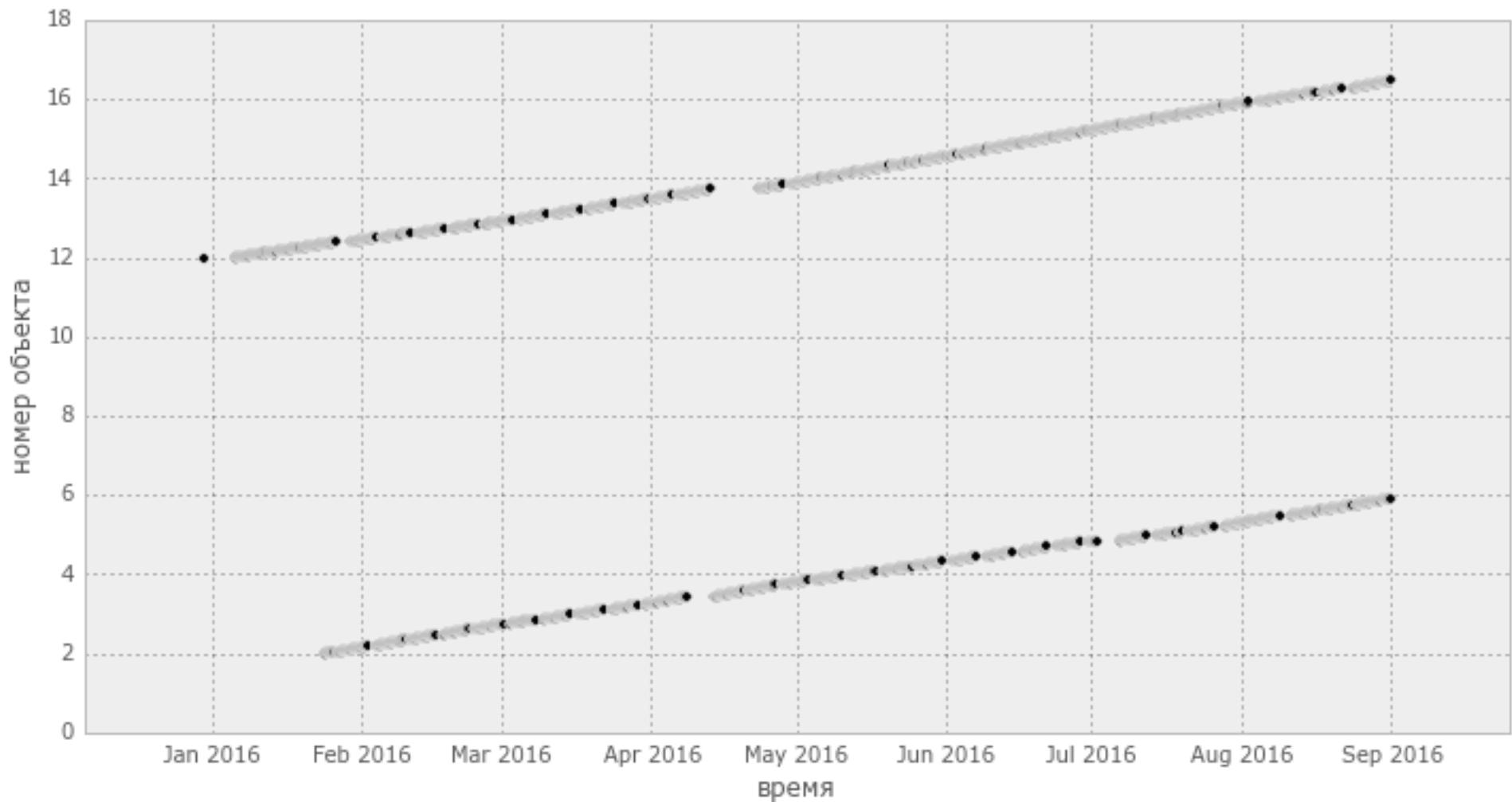
Пример dummy-визуализации



Сделайте график «id – время»:

- **простая проверка на монотонность**
- **видны «подозрительные периоды»**

Пример дутту-визуализации



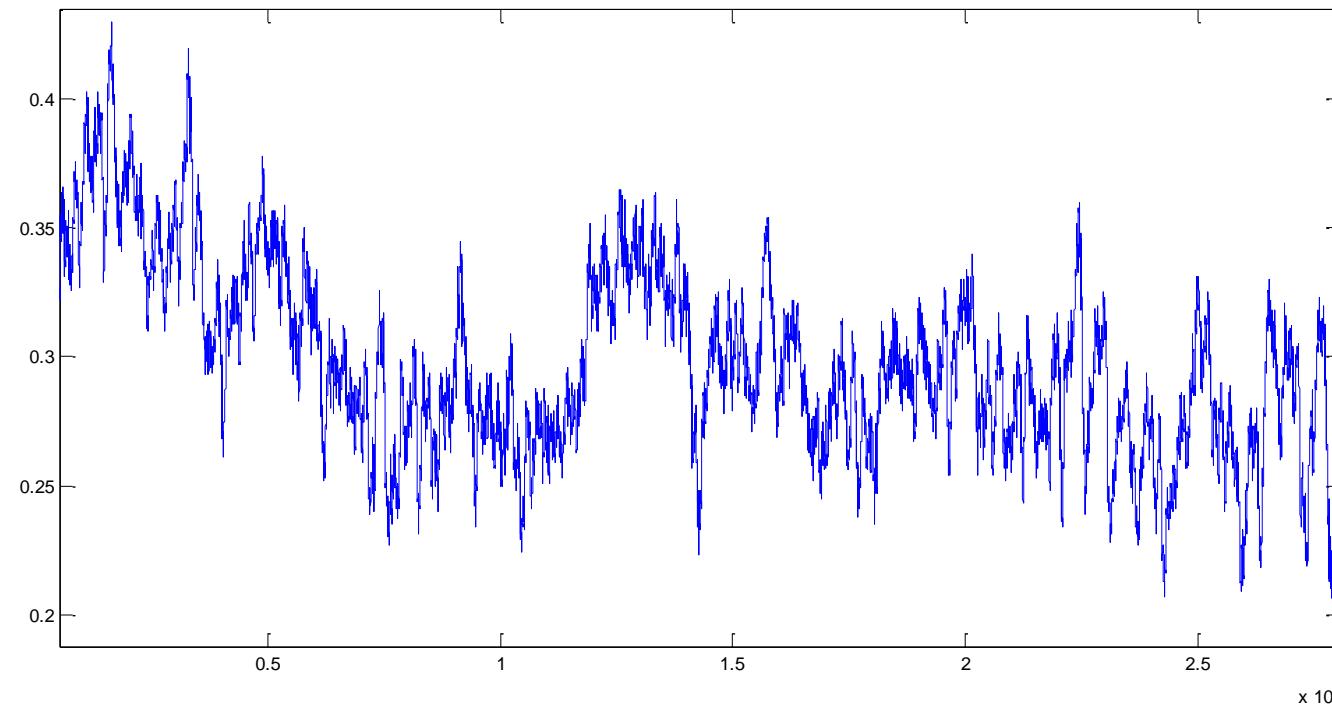
Случай из жизни: время – номер объекта

Видна двойная нумерация, периоды непоявления объектов

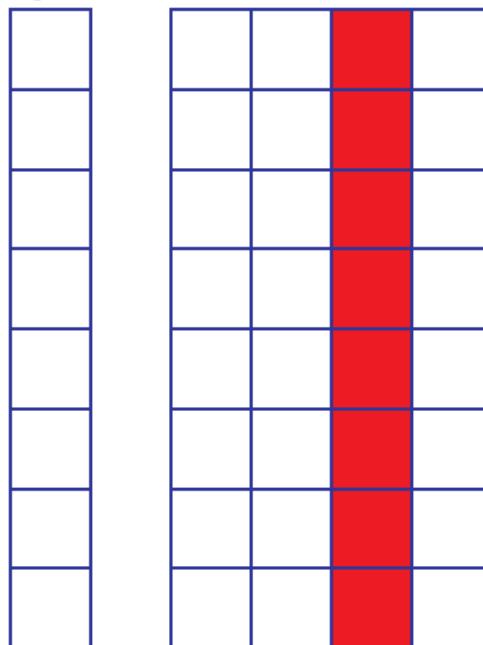
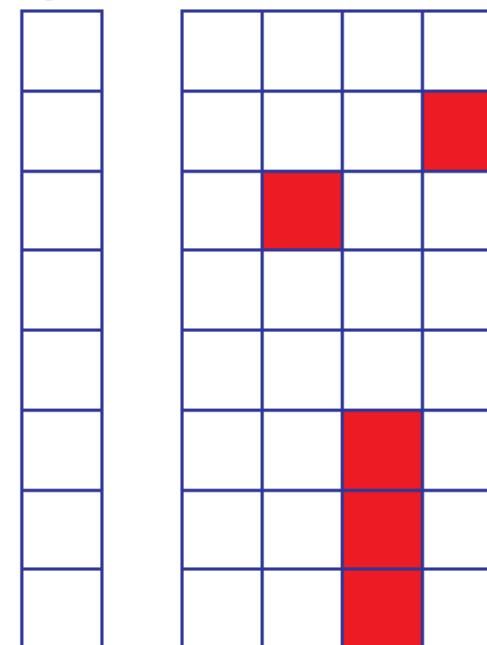
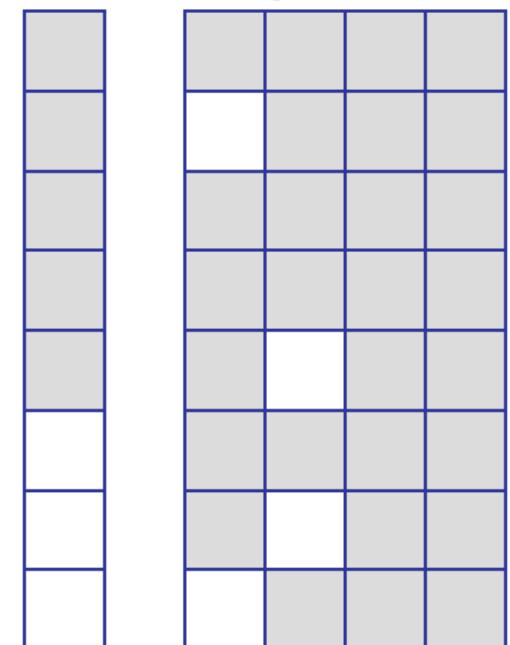
При раскраске по другим признакам видно больше!

Пример dummy-визуализации

Как меняется цель со временем



Применяется сглаживание окном

- шумовые признаки**удалить****Что есть в данных:**
- шумовые значения**причины:**
«ошибки из-за
невнимательности»,
«особые режимы»**метод:**
+dummy!!!**-пропуски:****причины:**
«нет значения»,
«не знаем значения»