курс «Машинное обучение»

# Искусство визуализации

## Часть 2. Одномерный анализ

Александр Дьяконов

17 марта 2021 года

# План

Зачем смотреть на данные

История визуализации и инфографики

Правила визуализации

**Одномерный анализ**

**Описательные статистики, их визуализации**

**Первичные действия при анализе признака**

**Визуализация отдельных признаков**

Многомерный анализ

Визуализация пары признаков

Визуализация «алгоритм» – «алгоритм/признак»

3D-визуализации

Dummy-визуализации

Игра «Что изображено?»

## Одномерный анализ (Univariate Analysis)

**– исследование отдельных признаков**

**здесь: «одномерные графики»**

**устанавливаем природу признаков**

**проверяем логичность признаков**

**для каждого признака**

- имя
- область значений
- распределение
- особенности (аномалии, пропуски и т.п.)
- устойчивость
- важность

**если что-то нарушается… пользуемся этим**

## Что просто визуализировать

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

# Описательные статистики – среднее

$$x_1 \le \ldots \le x_m$$

**Выборочное среднее**

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

**Мода (частое значение)**

$$\text{mode}(X) = \arg\max_x |\{i \in \{1,2,\ldots,m\} \mid x = x_i\}|$$

**Усечённое среднее**

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

**+ весовые схемы**

**+ сглаживание**

**mid-range (mid-extreme)**

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

**тоже одно из решений оптимизационных задач...**

**Медиана**

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

**midhinge**

$$\text{midhinge}(X) = \frac{q_{0.25} + q_{0.75}}{2}$$

# Описательные статистики – характерные элементы

**Минимум** $x_1$

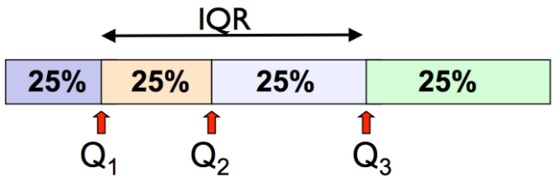
**Максимум** $x_m$

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

$$X = \{x_1, \ldots, x_m\}$$

**Квартили**

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



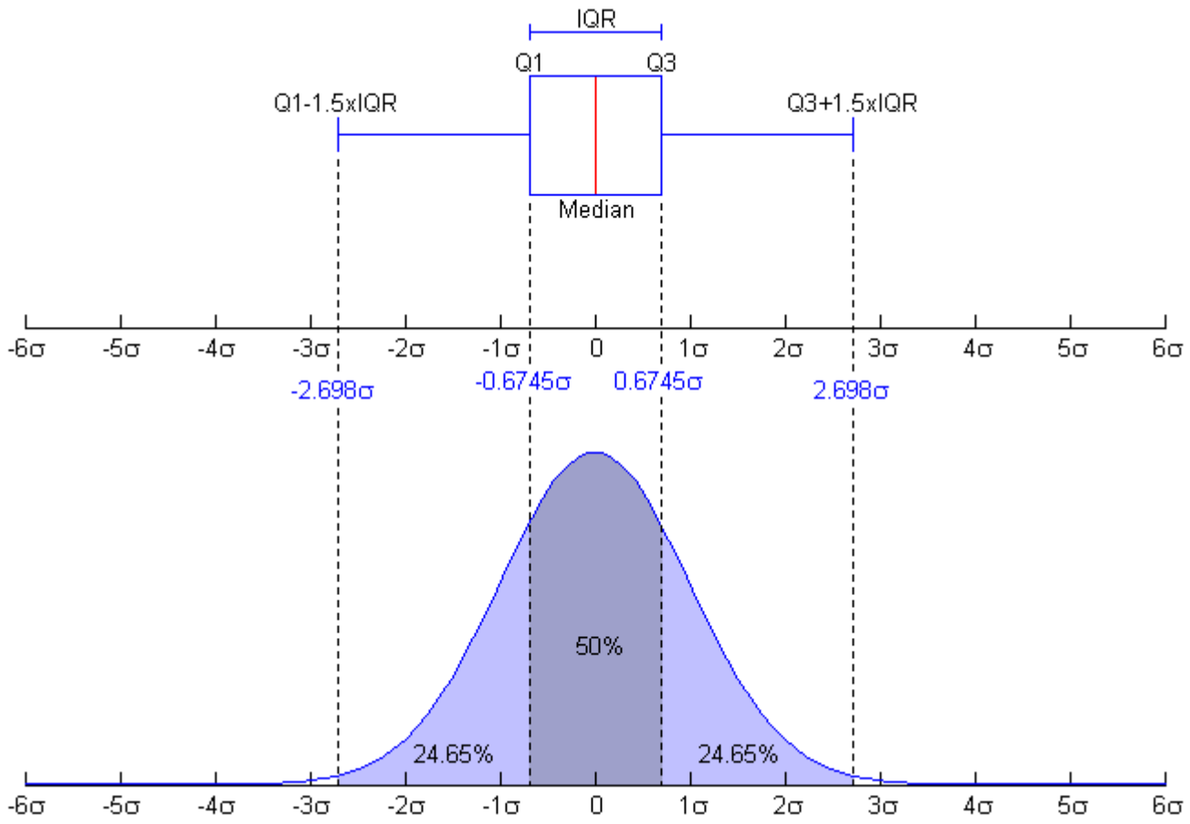

**Децили**

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

**Процентили**

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

# Описательные статистики – характерные элементы




**n-й элемент**

$$\text{nth}(X,k) = x_k$$

**м.б. для какого-то специального порядка**

**важная статистика!**

# Описательные статистики – разброс значений

## Среднее линейное (абсолютное) отклонение
## Mean Absolute Deviation

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

$\mathrm{mid}(X)$ – **любая формализация среднего**

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

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

# Описательные статистики – абсолютные вариации

**Чаще: стандартное отклонение**

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

**Размах**

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

**Дисперсия (рассеяние, разброс)**

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

**Median Absolute Deviation (MAD)**

$$\text{MAD}(X) = \text{median}(\{|\text{median}(X) - x_i|\}_{i=1}^{m})$$

**тоже обобщается на n-мерный случай**

**Среднее квартильное расстояние**

**Интерквартильный размах**

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

# Описательные статистики – абсолютные вариации

**Совет:**

$$\mathrm{mid}_2(\{|x_i - \mathrm{mid}_1(X)|\}_{i=1}^m)$$

$\mathrm{mid}_1$, $\mathrm{mid}_2$ – **любые формализации среднего**

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

## Максимальное абсолютное отклонение (Maximum Absolute Deviation)

$$\max(\{|x_i - \mathrm{mid}(X)|\}_{i=1}^m)$$

**Именно это оптимизирует** $\mathrm{mid\text{-}range}$

**Тут могут быть любые функции!**

# 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 – центральные моменты

$$\mathbf{E}[(X - \mathbf{E}X)^k]$$

**1, 0, дисперсия, …**

## Описательные статистики – моменты

$$\mathbf{E}[X^k]$$

# Описательные статистики – стандартизованные моменты
# Standardized moments

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

$k=1$ **0**

$k=2$ **1**

**Асимметрия – skewness**

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

**Эксцесса (островершинность) – kurtosis**

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

**инвариантны относительно изменения масштаба**

# Описательные статистики – другое

## Стандартная ошибка среднего

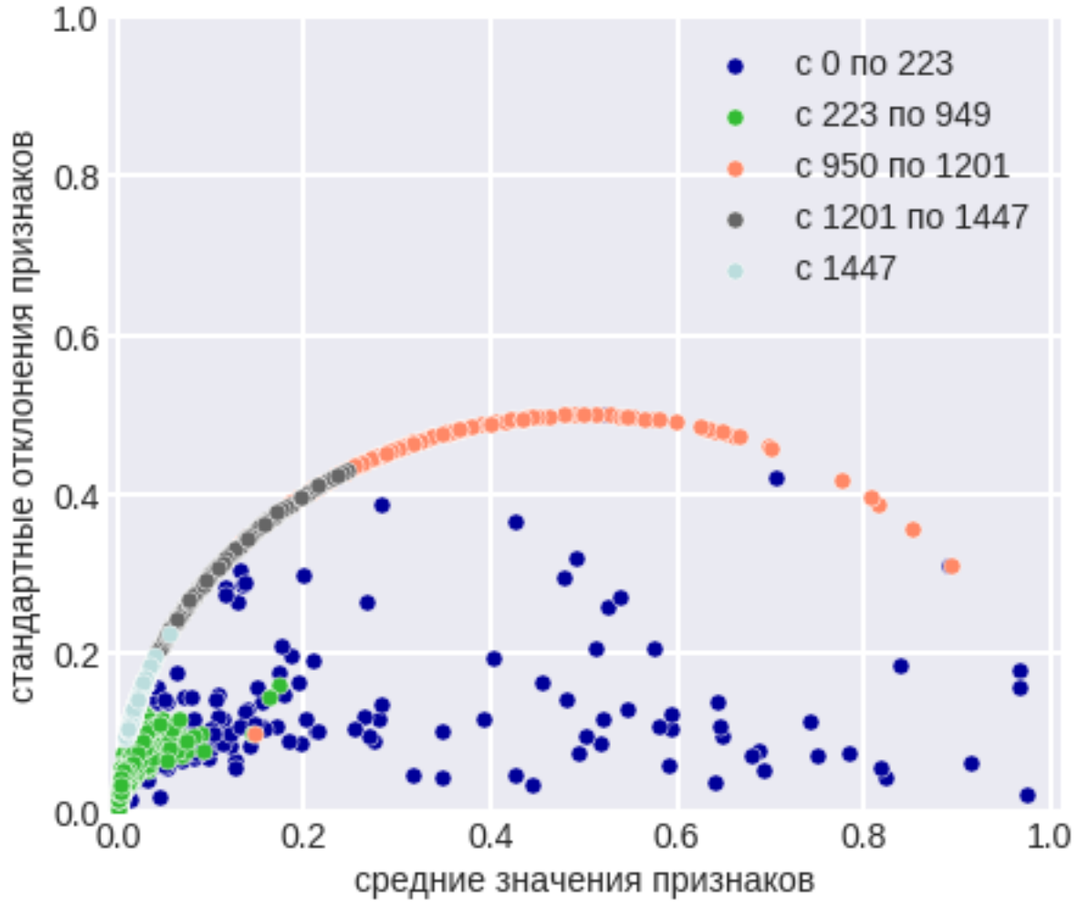
$$\frac{\text{std}(X)}{\sqrt{m}}$$

# Визуализация описательных статистик: задача Biological Response



**Чётко видны группы**

# Визуализация описательных статистик: задача Biological Response




**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, 0, 1, 0, 0]**

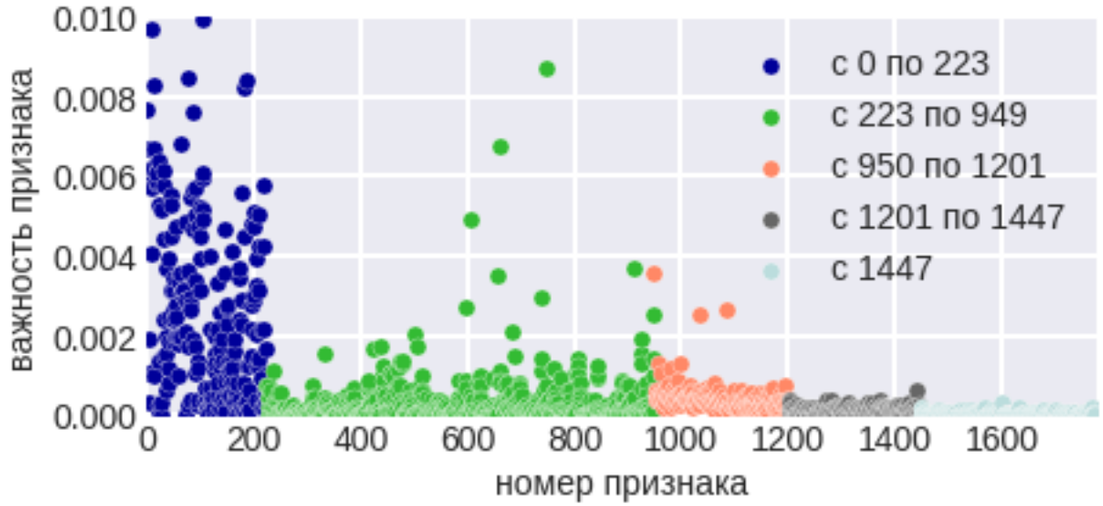
$$\text{mean}(\{x_i\}_{i=1}^m) = \frac{1}{m}\sum_{l=1}^{m} x_i \equiv p$$

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

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

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

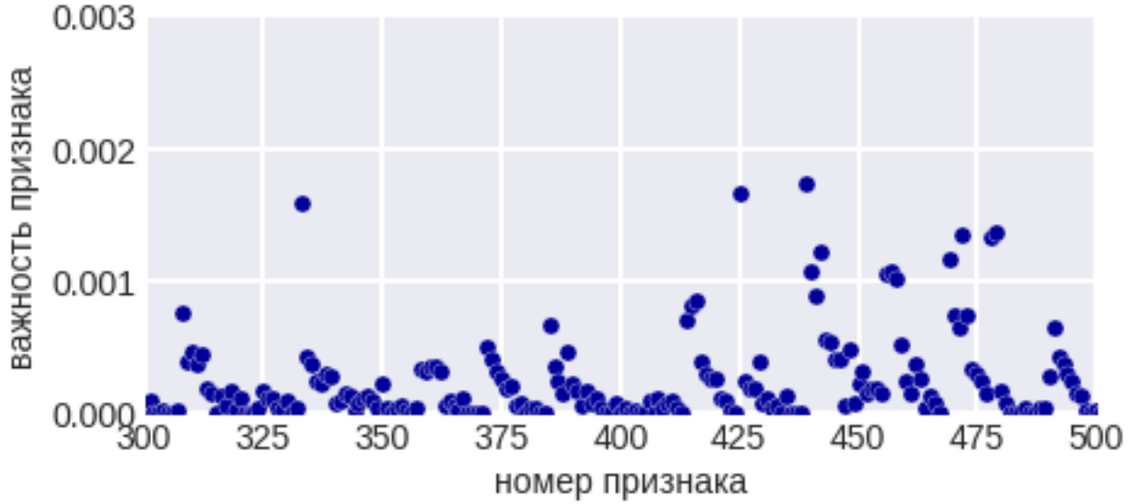
# Визуализация важностей признаков: задача Biological Response




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

# Визуализация важностей признаков: задача Biological Response

## Увеличение картинки




**Есть подгруппы признаков!**

**Меняйте масштаб!**

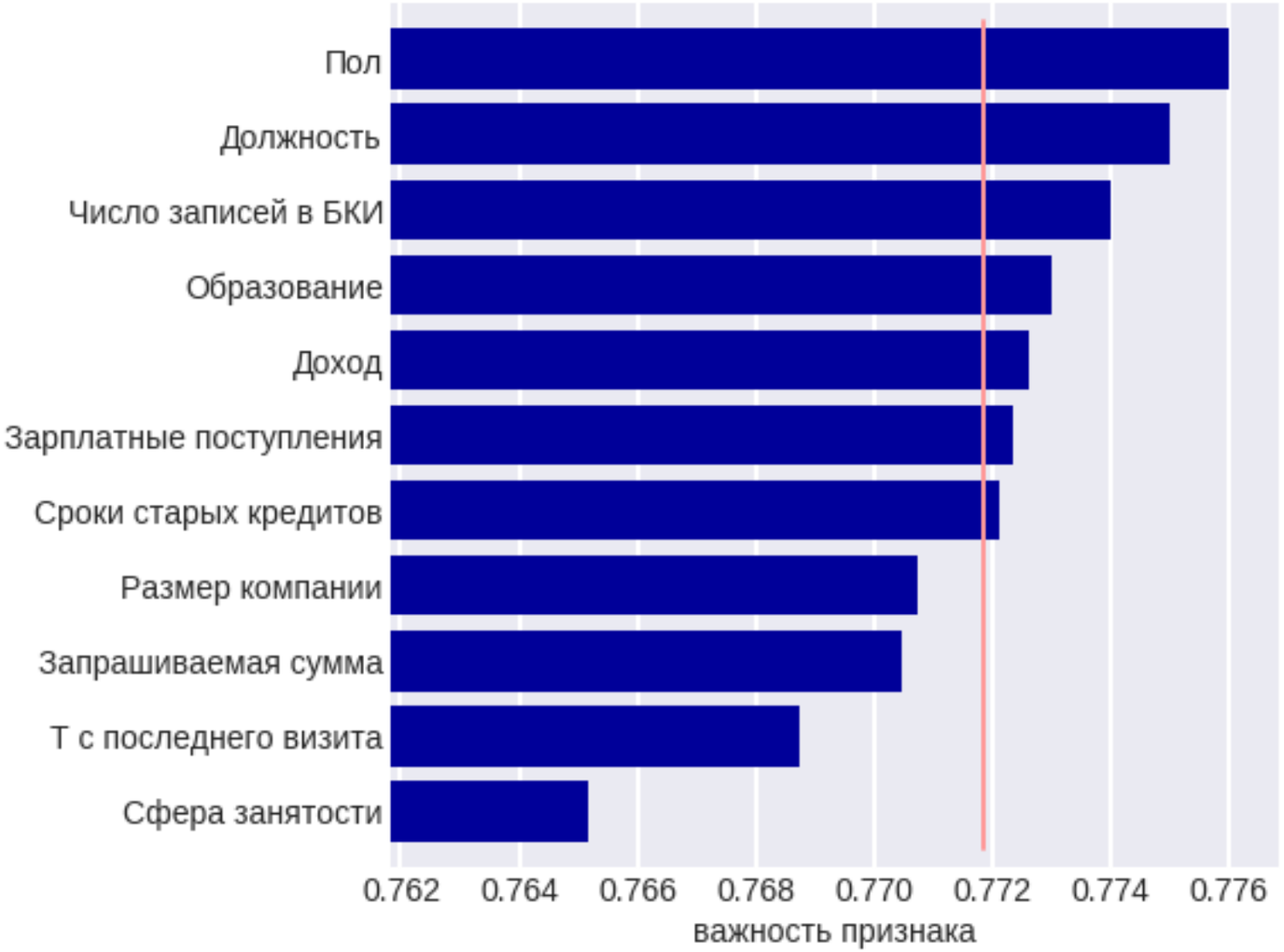
# Аналогично – исследование сложности «классификации» объектов

## Исследование частей выборки (фолдов)


**Подозрительная унимодальная зависимость!**

**Что значит?**

# Как правильно показывать важности признаков




**Сортировка, среднее значение, вертикальная ориентация**

# Правило столбцовых диаграмм

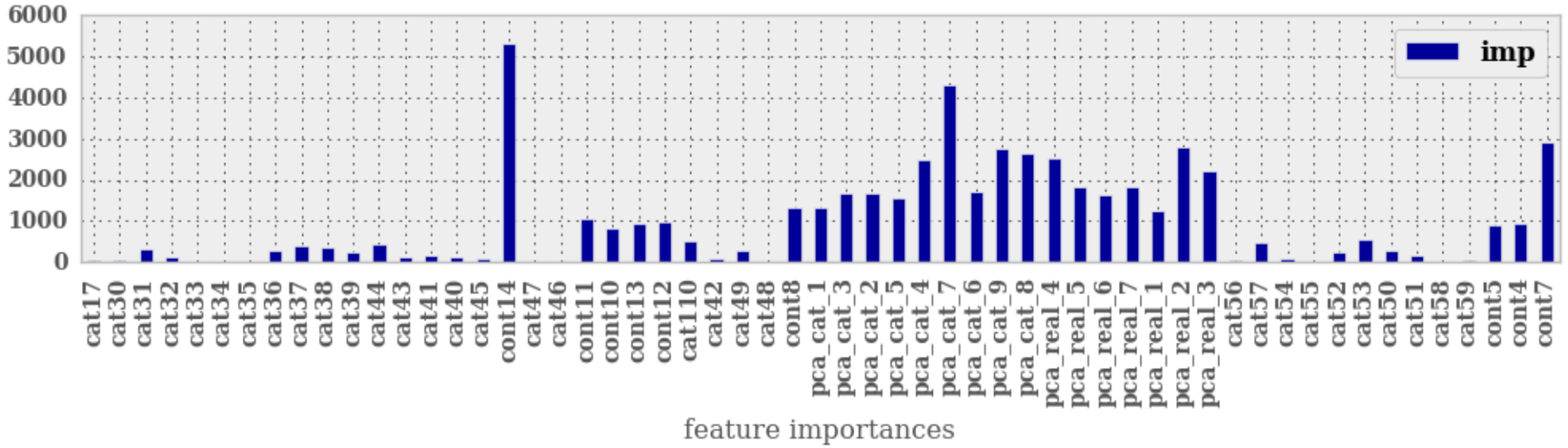


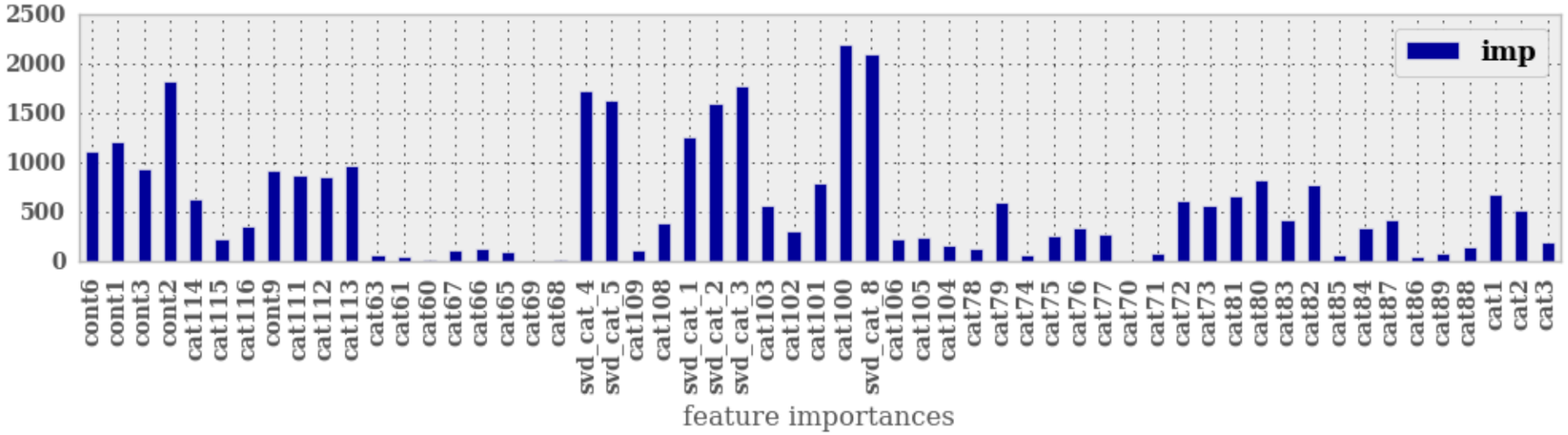
**Правило:**

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

**Про важности в отдельной лекции**

# Важности признаков



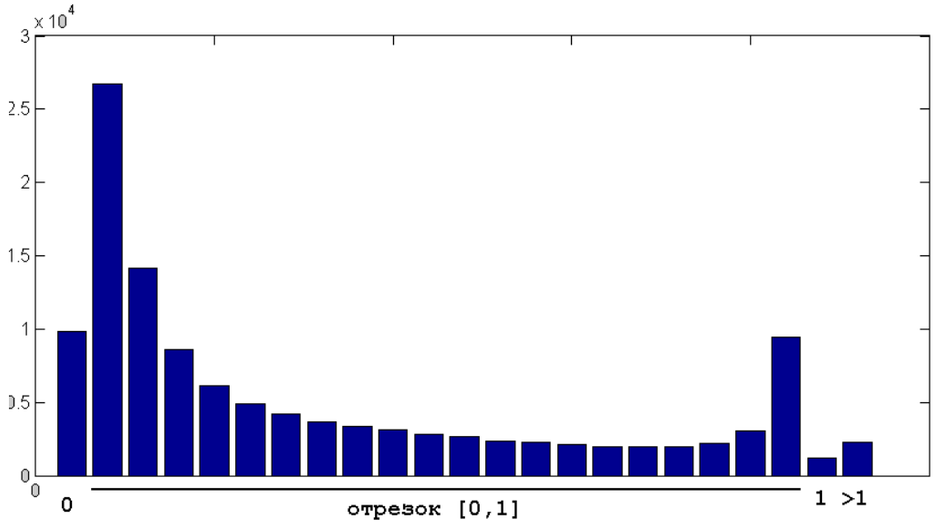

## Придумываем признаки и анализируем «AllState»

# 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)))
```

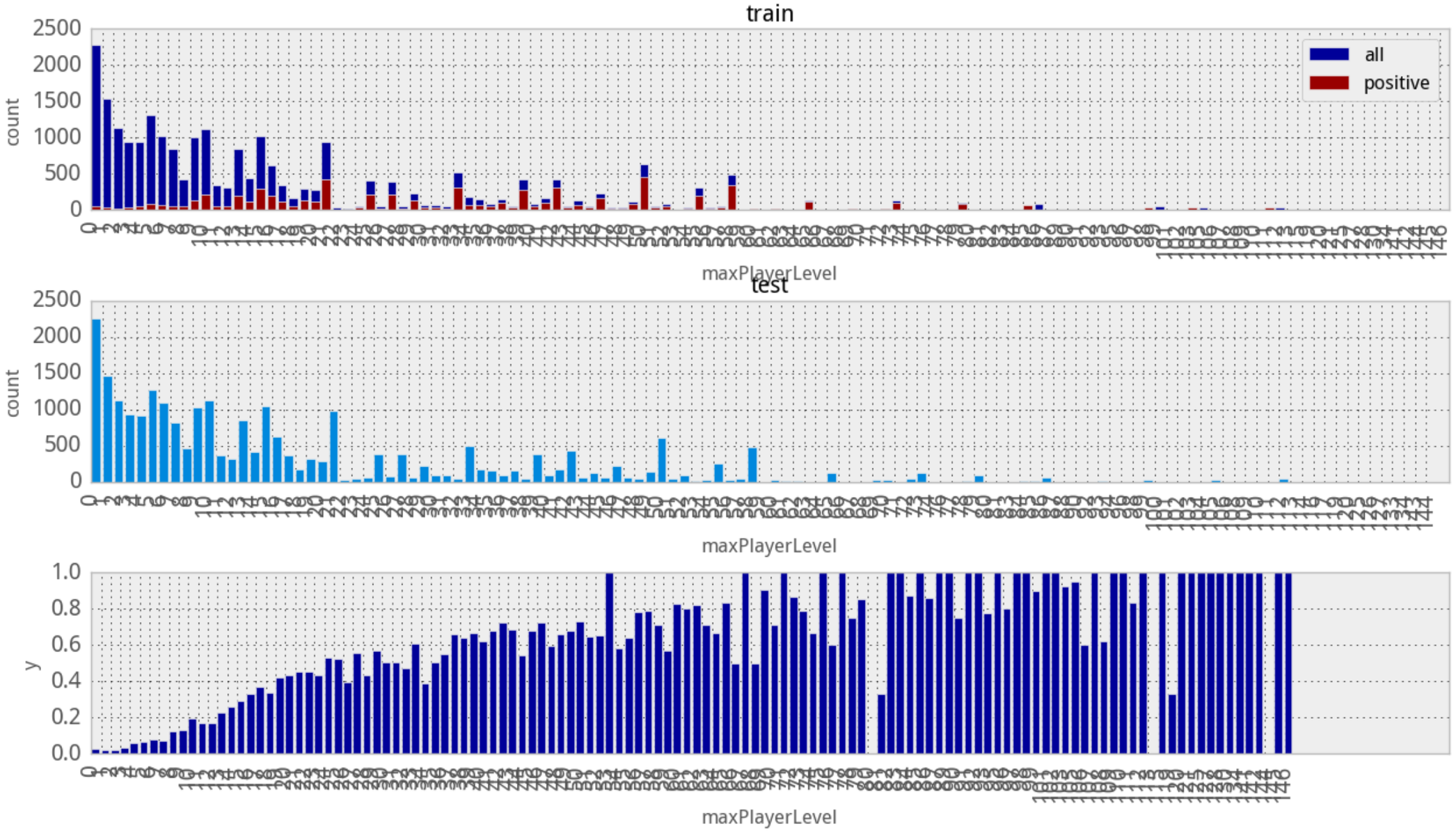
```
                    Класс          4       2.20       0.97 [1 2 3 4]
                    Номер       8404    7442.45     269.63 [5001 5002 ...]
                  Вес, т         124      38.27       7.30 [ 41.1  44.4  ...]
                  Начало       8404                  time
           Количество,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

# Что надо сразу выяснить про признак

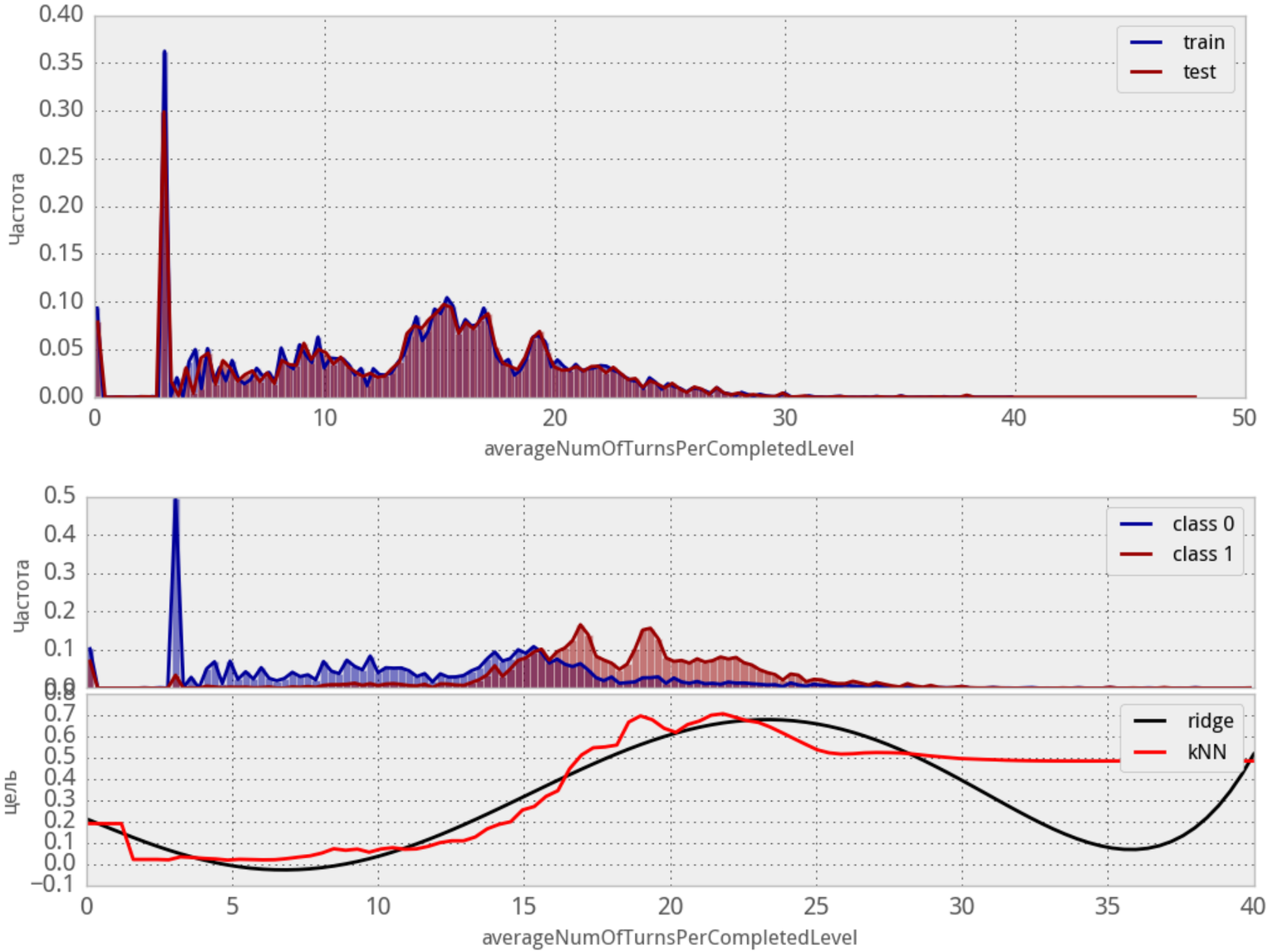
- распределение значений признака
- распределение обучение / тест
- распределение целевой переменной (ex: класс 0 / 1)

- такие же вопросы для пропусков, выбросов

# Что надо сразу выяснить про признак

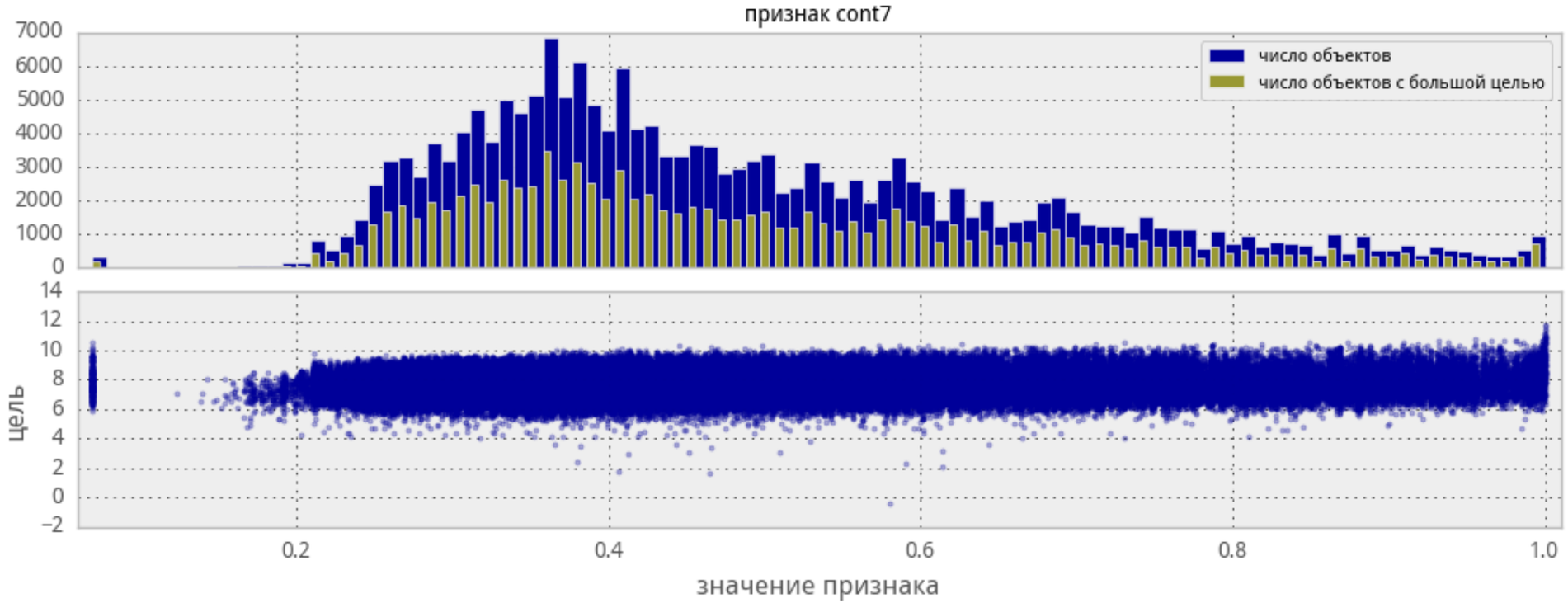


# Что надо сразу выяснить про признак



# Что надо сразу выяснить про признак

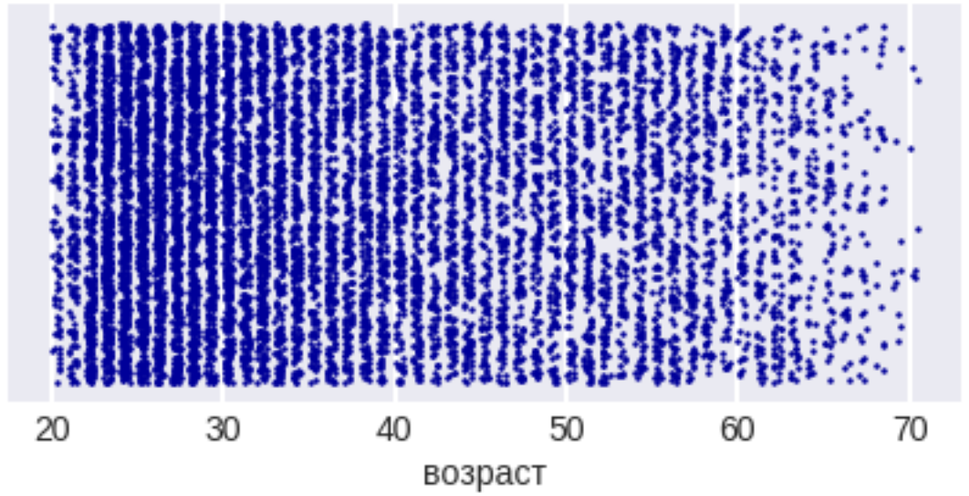
## «AllState»

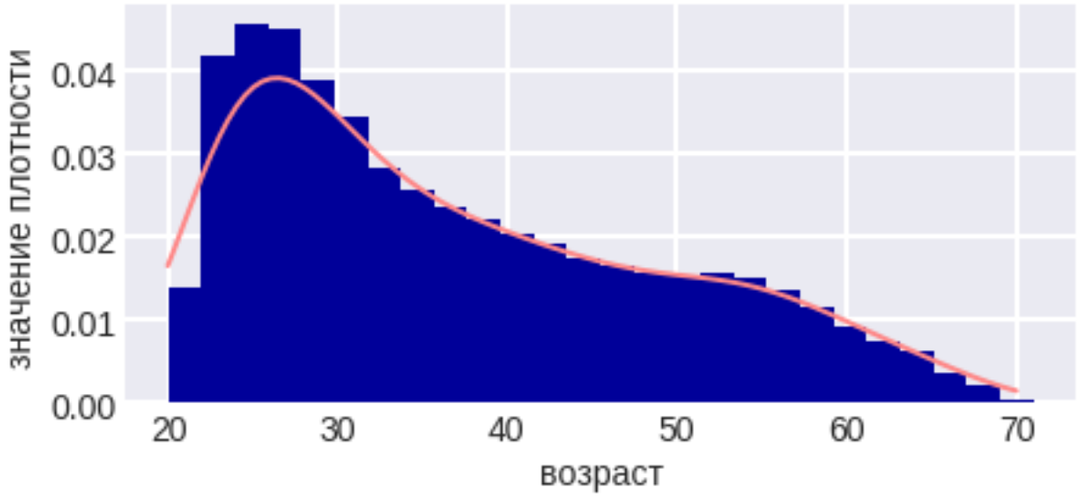



**Вверху – гистограмма распределения по значениям признака**
**Отдельно по объектам с большим значением целевого признака**

**Внизу – диаграмма рассеивания «признак – цель»**

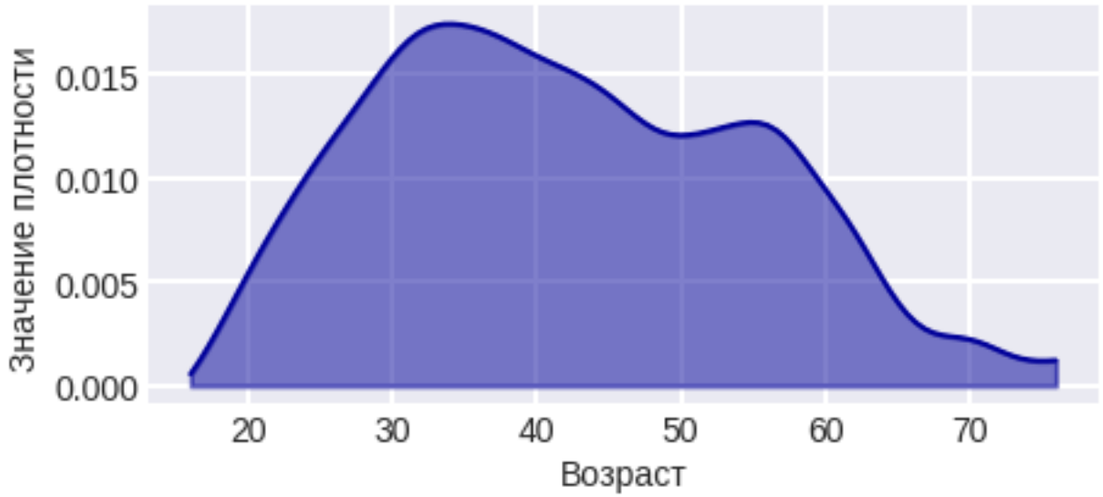
# Визуализация отдельных признаков




**Гистограммы предпочтительнее плотностей**

# ЗАДАЧА «М-магазин»

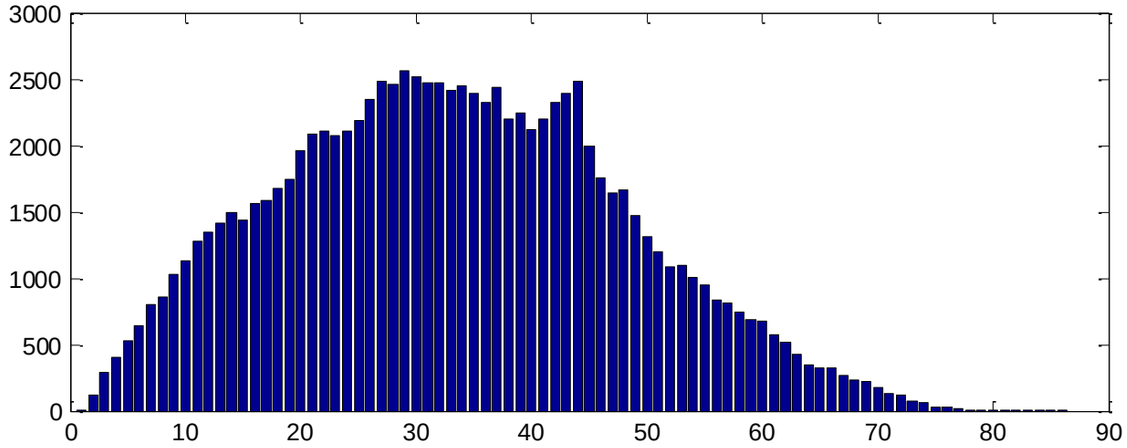



**Распределение возраста покупателей**

**Так обычно выглядит распределение!**

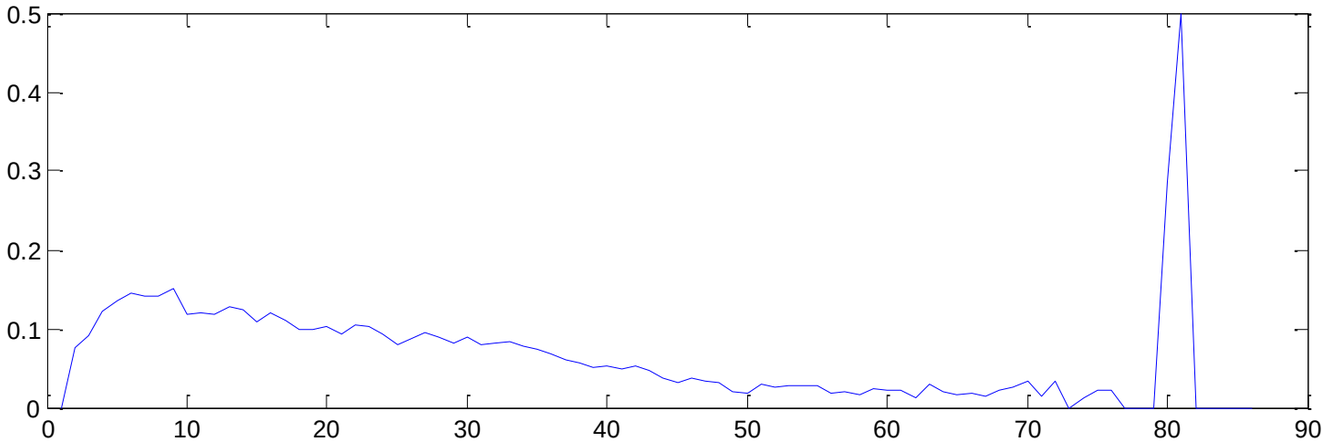
**Почему два горба?**

# ЗАДАЧА «ТКС»



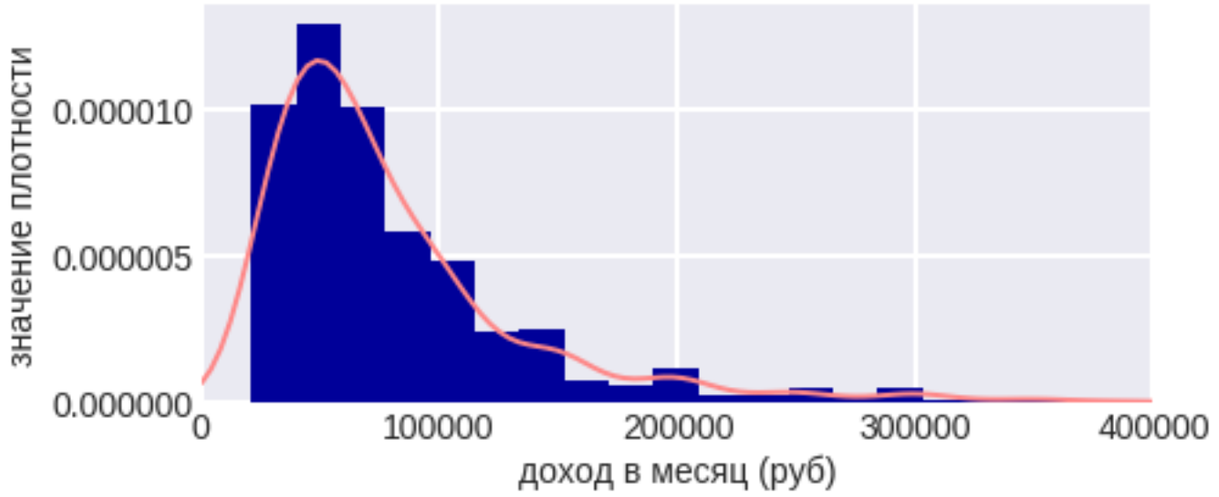

**Распределение по возрасту**

**Что значит?**




**Отношение плотностей – есть явный выброс!**

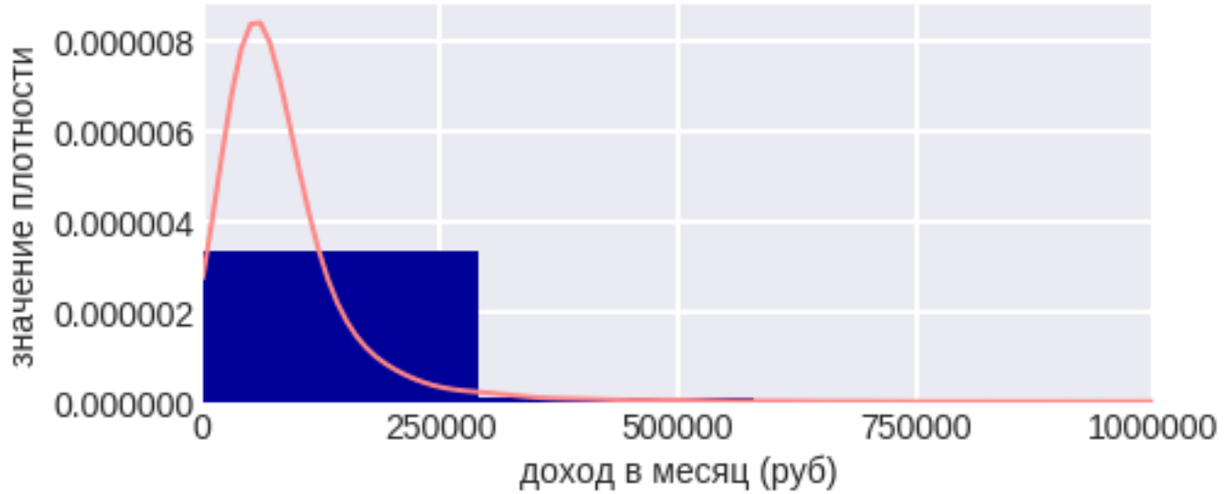
# Проблемы визуализаторов – параметры по умолчанию





**увеличили число бинов**

# Проблемы визуализаторов – выбросы



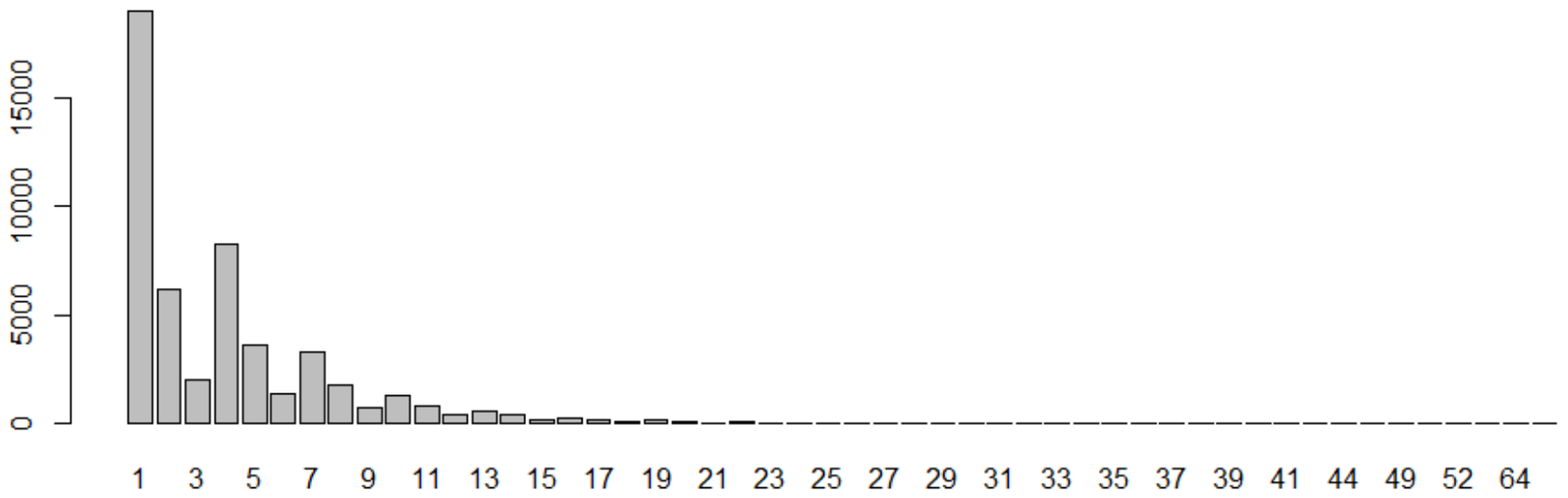

## Что будет если не устранять выбросы…

```
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»


**Из-за правильной визуализации**

**немонотонная зависимость**

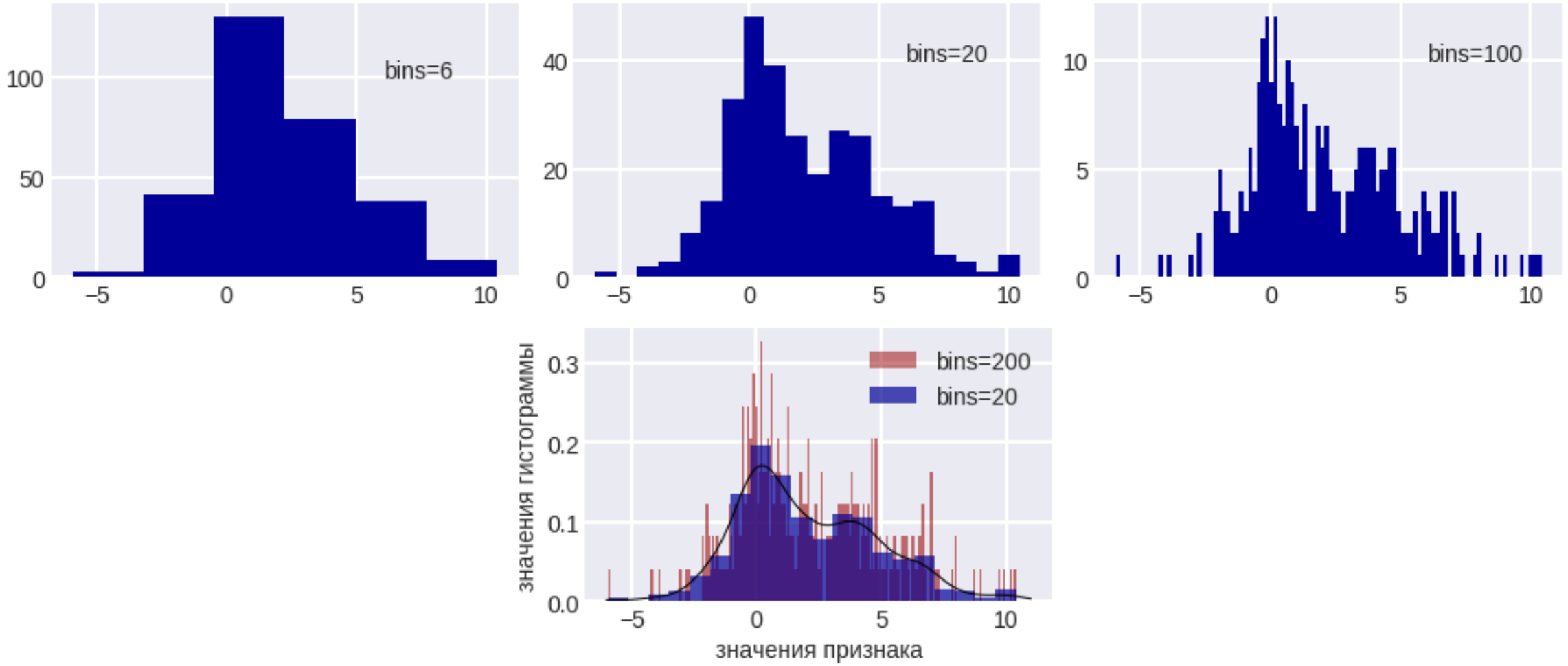
**паттерны – «тройки»**

**Выбирать:**

**число бинов**

**ширина столбцов**

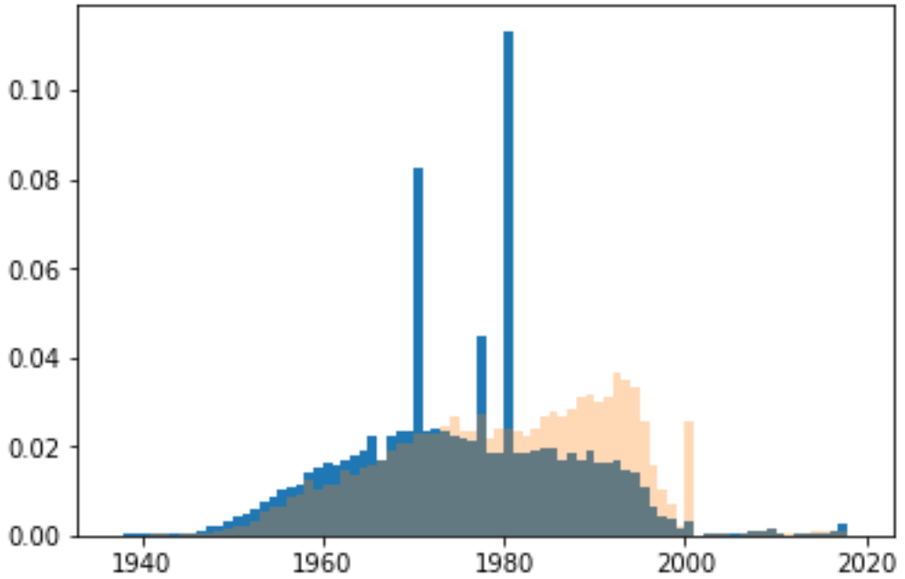
# Построение гистограммы




**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

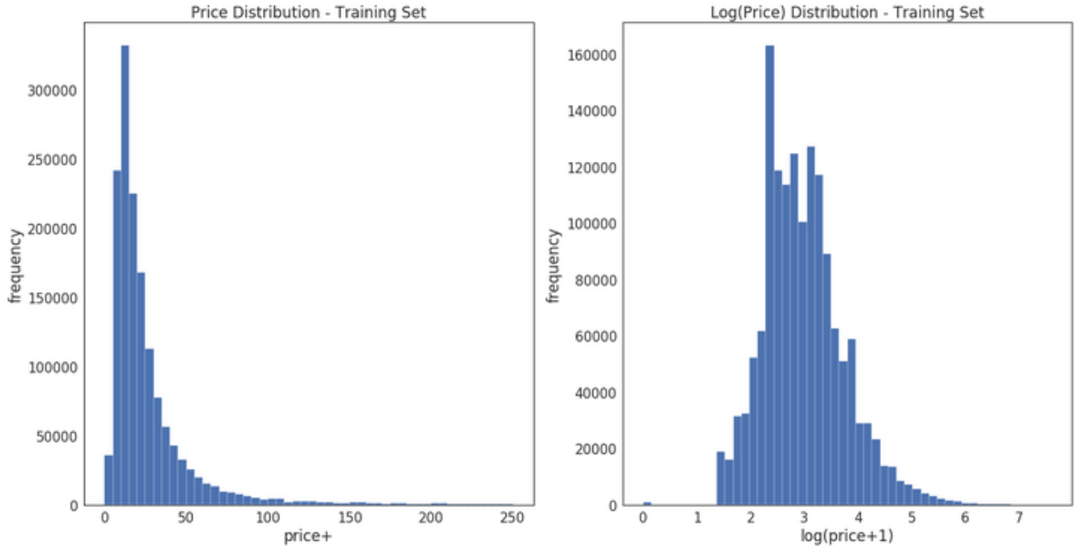
**значения по умолчанию $\Rightarrow$ точная дата неизвестна**

**при этом пол «Ж» $\Rightarrow$ тоже неверно**

**Стоит ли доверять другой информации?**

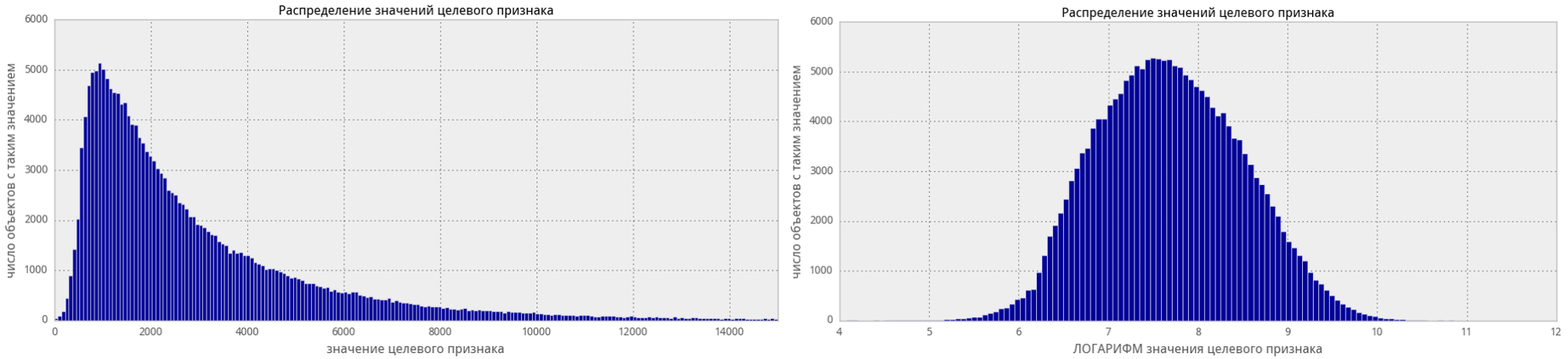
# Выводы о признаках

## Использование визуализации для выбора трансформации



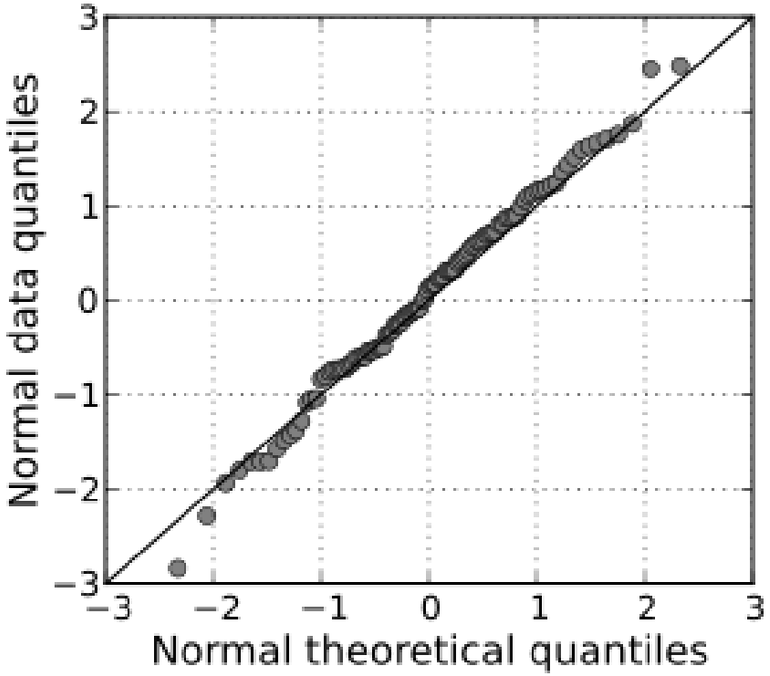

https://www.kaggle.com/thykhuely/mercari-interactive-eda-topic-modelling

## «AllState»



# Анализ распределения

## Q–Q (quantile-quantile) plot




https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

# Визуализация отдельных признаков

**Приёмы**

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

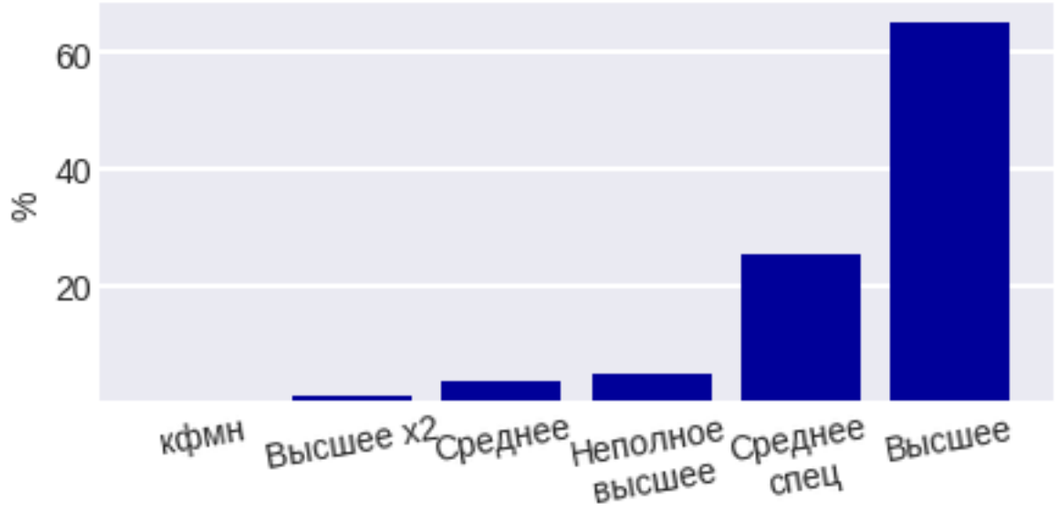
**Зачем**

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

**Сравнение:**

- **при разных значениях целевого**
- **на обучении и контроле**

# Визуализация категориальных признаков

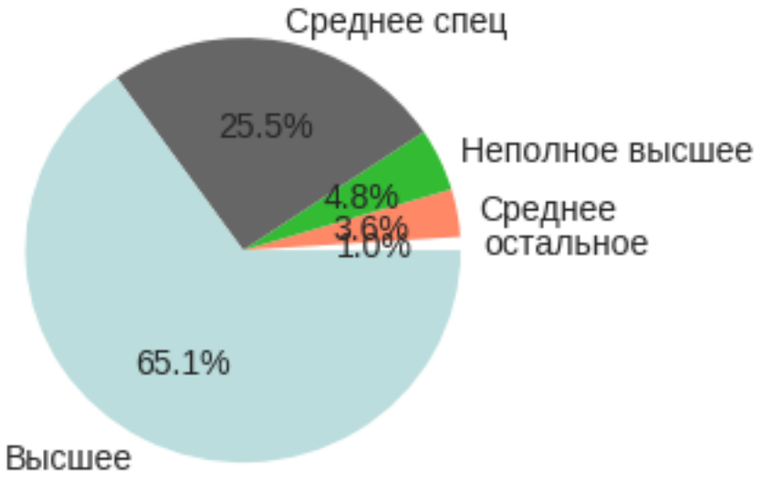


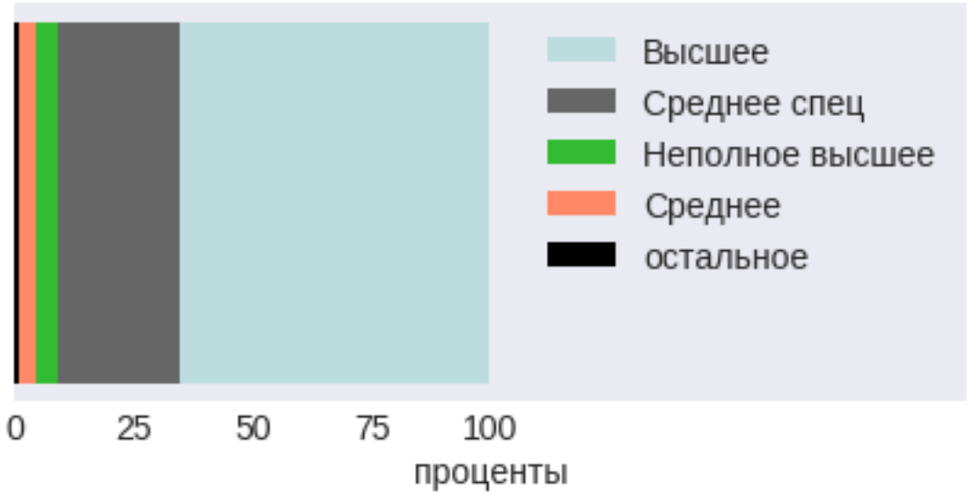

**не видно мелкие категории**

**категорий может быть много**

**Как быть?**

# Визуализация категориальных признаков



# Визуализация категориальных признаков

**Не использовать 3D-эффекты**

**Мелкие категории → «остальное»**

**Площадь всех категорий = 100%**

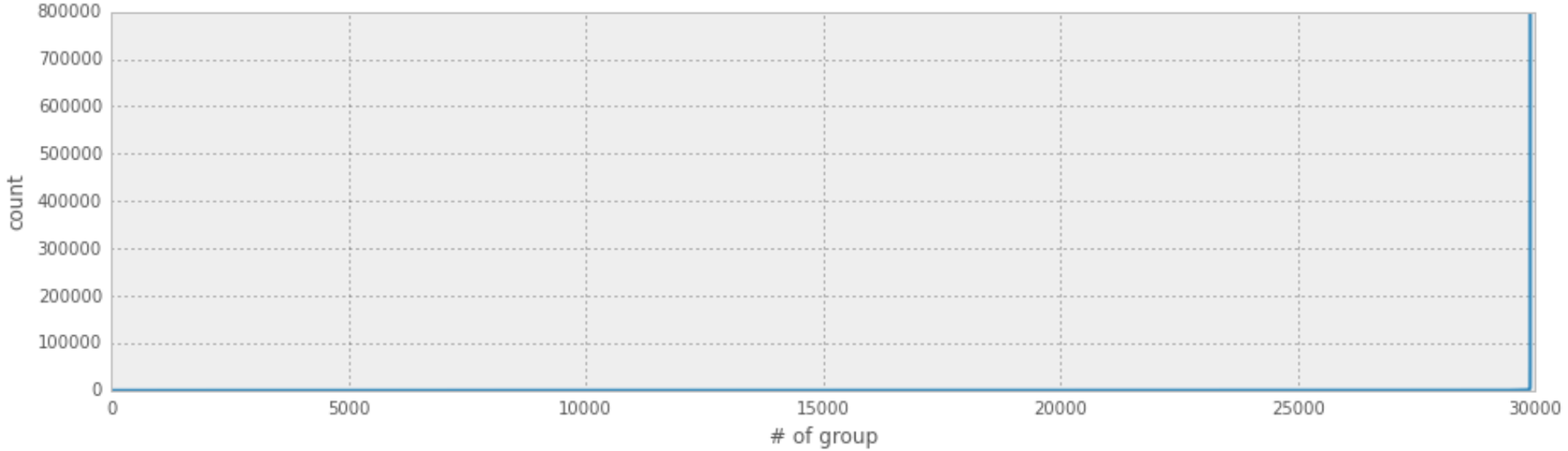
**Диаграмма-пирог – не рекомендуется**

**Когда информации для визуализации мало – таблицы!**

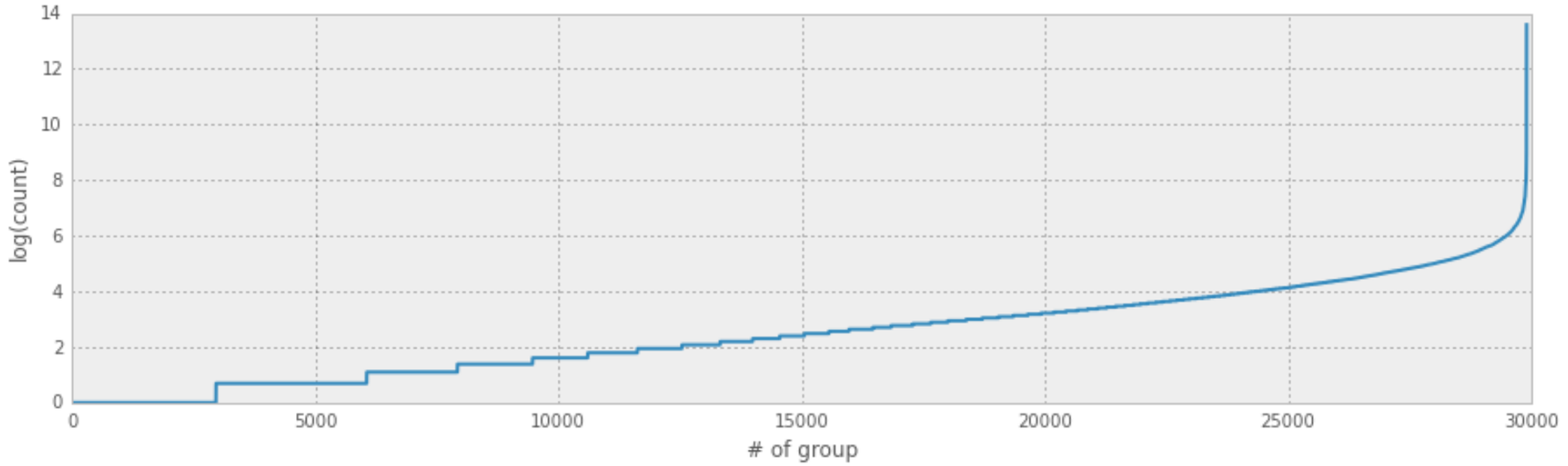
| Образование | % |
|---|---|
| Высшее | 65.1 |
| Среднее спец | 25.5 |
| Неполное высшее | 4.8 |
| Среднее | 3.6 |
| Высшее х2 | 0.8 |
| кфмн | 0.2 |

**Можно ещё логарифмировать…**

# Зачем ещё нужно логарифмирование



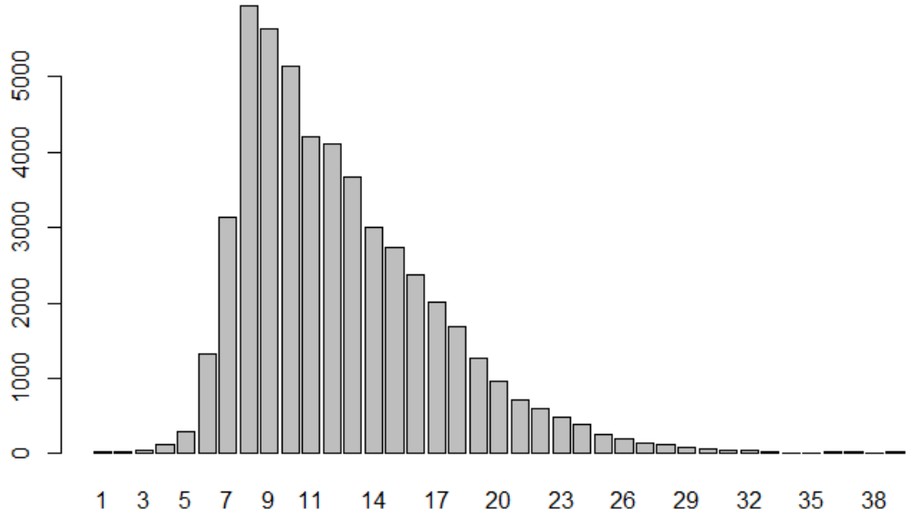

**число представителей одной из ~30000 групп в выборке**




**логарифм этого числа**

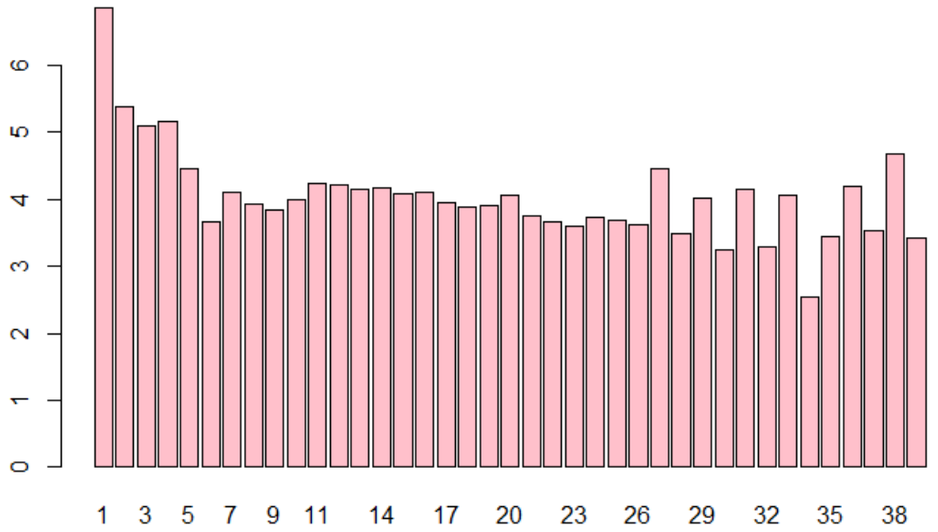
# Распределения на признаках – природа признаков

## 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»

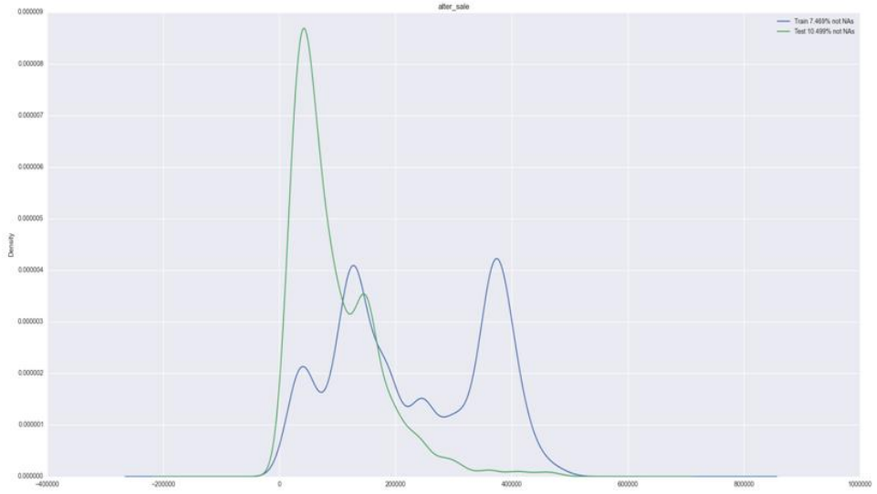
| | mean | count |
|---|---|---|
| | cat101 | |
| A | 2454.139844 | 106721 |
| B | 1292.020000 | 3 |
| C | 2778.283638 | 16971 |
| D | 2812.990306 | 17171 |
| E | 4458.574286 | 7 |
| F | 3560.151861 | 10139 |
| G | 3450.680947 | 10944 |
| H | 1320.720000 | 1 |
| I | 4590.935254 | 6690 |
| J | 4603.863790 | 7259 |
| K | 3240.165000 | 2 |
| L | 5321.419556 | 3173 |
| M | 5540.292766 | 3669 |
| N | 2192.720000 | 1 |
| O | 6870.387172 | 2493 |
| Q | 7057.470264 | 2762 |
| R | 8564.376594 | 138 |
| S | 8993.138439 | 173 |
| U | 15972.490000 | 1 |

| | cat107 | |
|---|---|---|
| A | 3259.510800 | 75 |
| B | 19845.900000 | 2 |
| C | 2076.430704 | 213 |
| D | 2636.230164 | 3225 |
| E | 2871.429175 | 12521 |
| F | 3072.621189 | 47310 |
| G | 3149.791915 | 28560 |
| H | 3124.043153 | 23461 |
| I | 2913.988215 | 20066 |
| J | 3084.531566 | 22405 |
| K | 2946.549609 | 20236 |
| L | 3003.206170 | 6976 |
| M | 3074.337929 | 2067 |
| N | 3053.982033 | 797 |
| O | 2950.613520 | 125 |
| P | 3138.672300 | 100 |
| Q | 2985.114143 | 140 |
| R | 3063.068000 | 5 |
| S | 5553.495000 | 2 |
| U | 3546.898438 | 32 |

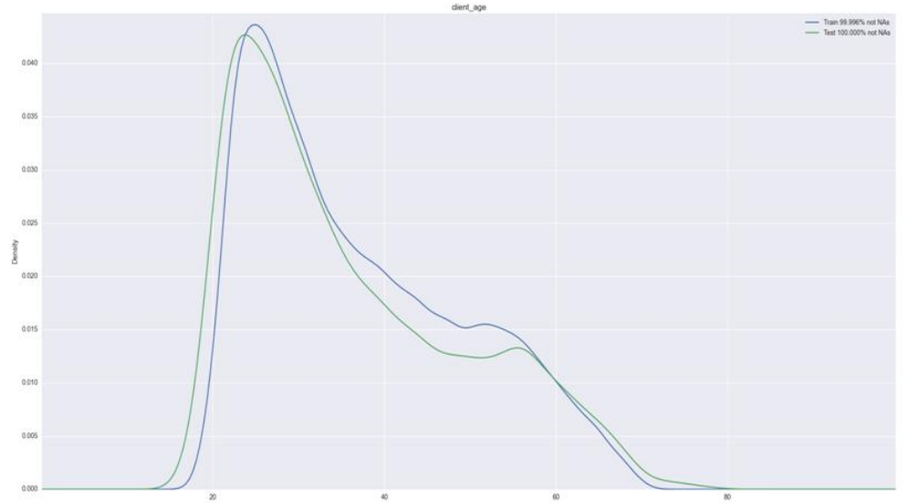
# Как распределение меняется при переходе к контролю

**смотреть как меняются распределения**

**обучение – контроль**




**Есть существенные изменения**




**Нет изменений**

**История про о-трэвел и волшебный признак.**

# Итог

**Гистограммы очень хороши**

- быстро оценить форму распределения
- придумать деформацию

но надо настраивать вручную (впрочем, любую визуализацию)

**Есть много описательных статистик**

хороши как признаки

**Смотреть по признакам**

распределения, распределения обучение / тест, распределения целевой переменной, аномальности в распределении, пропуски, естественность порядка значений

**Приёмы:**

**деформация признака** (чаще логарифмирование)

**масштабирование**

# Итог

- **не используйте сложных средств визуализации**

**Досконально понимать, как происходит сама визуализация!**

- **не используйте параметров по умолчанию при визуализации**

**«Посмотреть на данные» — это тоже процедура, которая нуждается в обучении, т.е.**

- **настройке параметров**
- **м.б. очистка данных от выбросов**
- **м.б. изменение шкалы (например, логарифмирование)**