курс «Машинное обучение»

Искусство визуализации

Часть 2. Одномерный и многомерный анализ

Александр Дьяконов

План

Одномерный анализ

Описательные статистики, их визуализации

Первичные действия при анализе признака

Визуализация отдельных признаков

Многомерный анализ

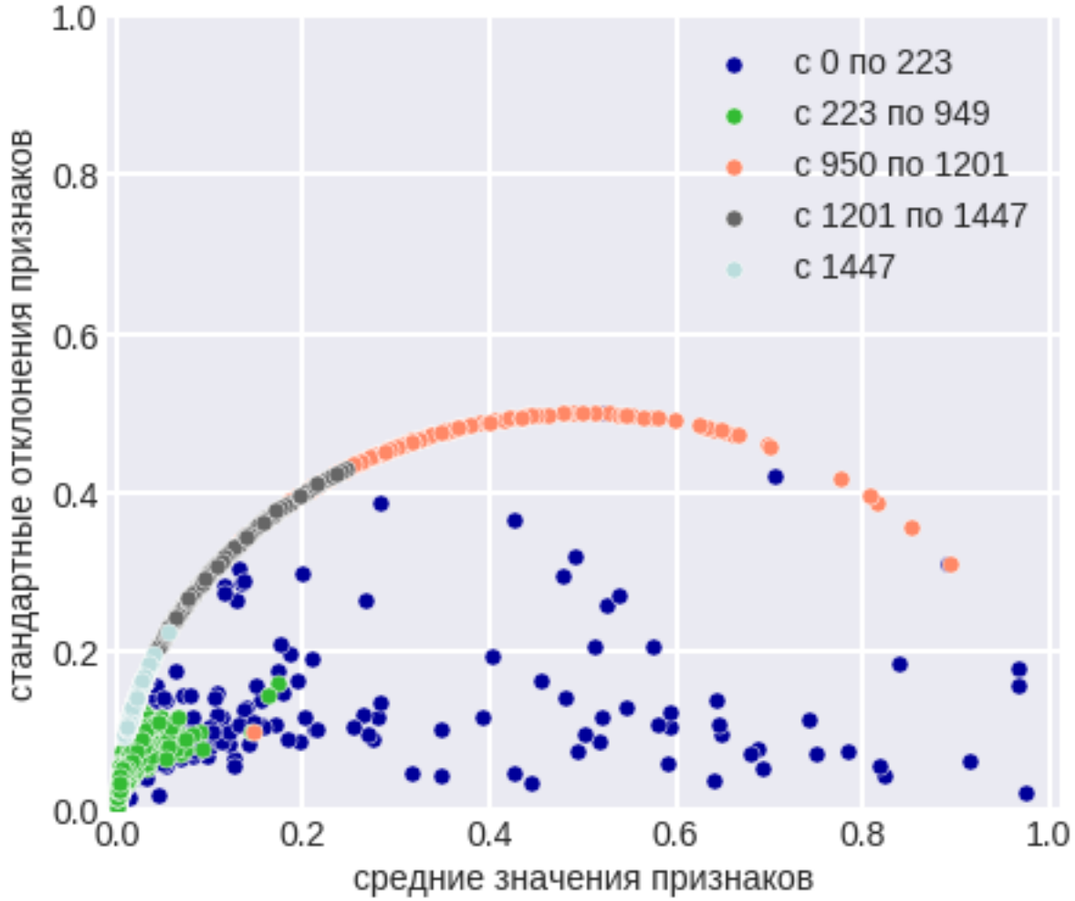
Визуализация пары признаков

Визуализация «алгоритм» – «алгоритм/признак»

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]

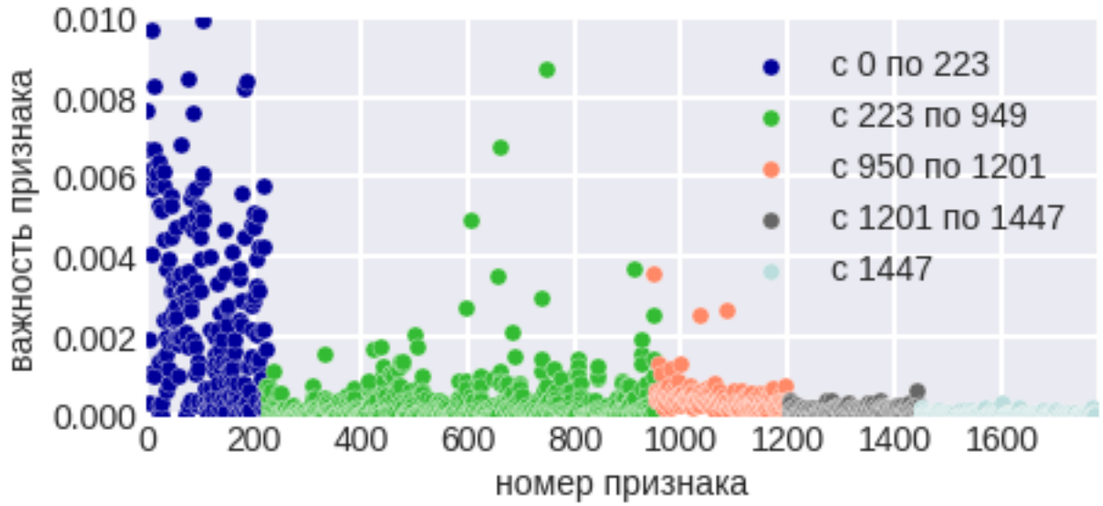
$$\text{mean}(\{x_i\}_{i=1}^m) = \frac{1}{m}\sum_{l=1}^{m} x_i \equiv p$$

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

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

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

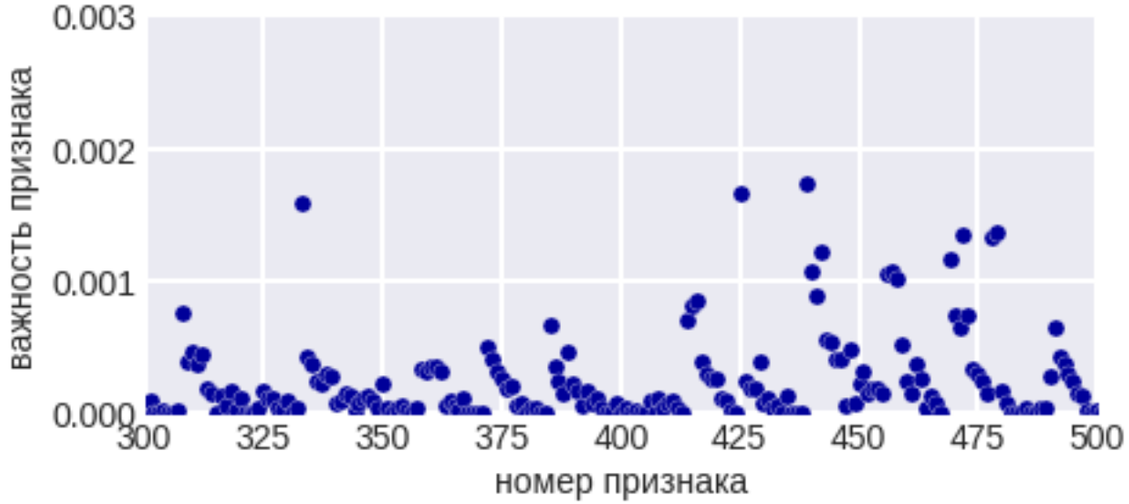
Визуализация важностей признаков: задача Biological Response



Потом: целые группы признаков можно удалять без существенной потери качества

Визуализация важностей признаков: задача Biological Response

Увеличение картинки



Есть подгруппы признаков!

Меняйте масштаб!

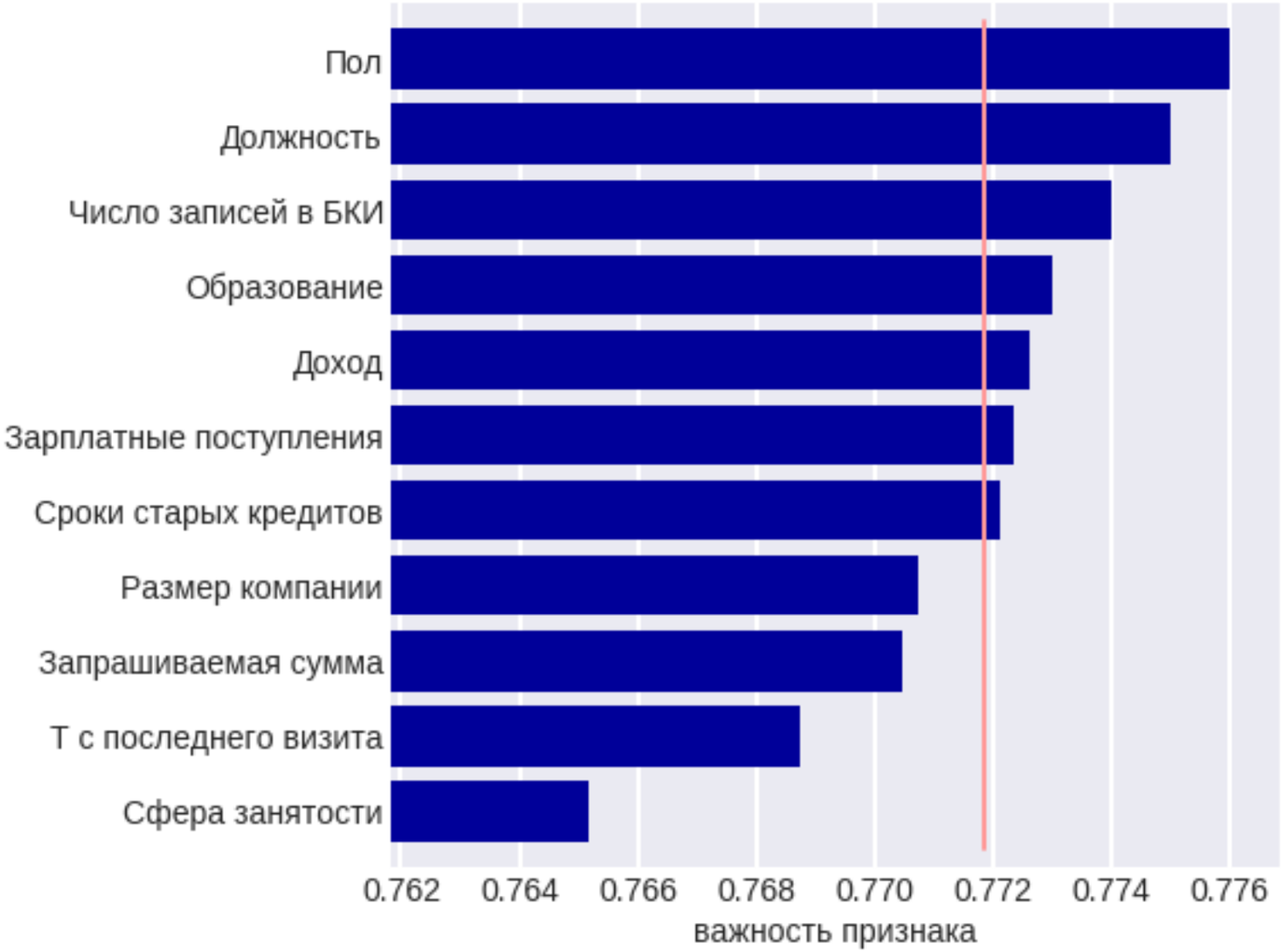
Аналогично – исследование сложности «классификации» объектов

Исследование частей выборки (фолдов)

Подозрительная унимодальная зависимость!

Что значит?

Как правильно показывать важности признаков



Сортировка, среднее значение, вертикальная ориентация

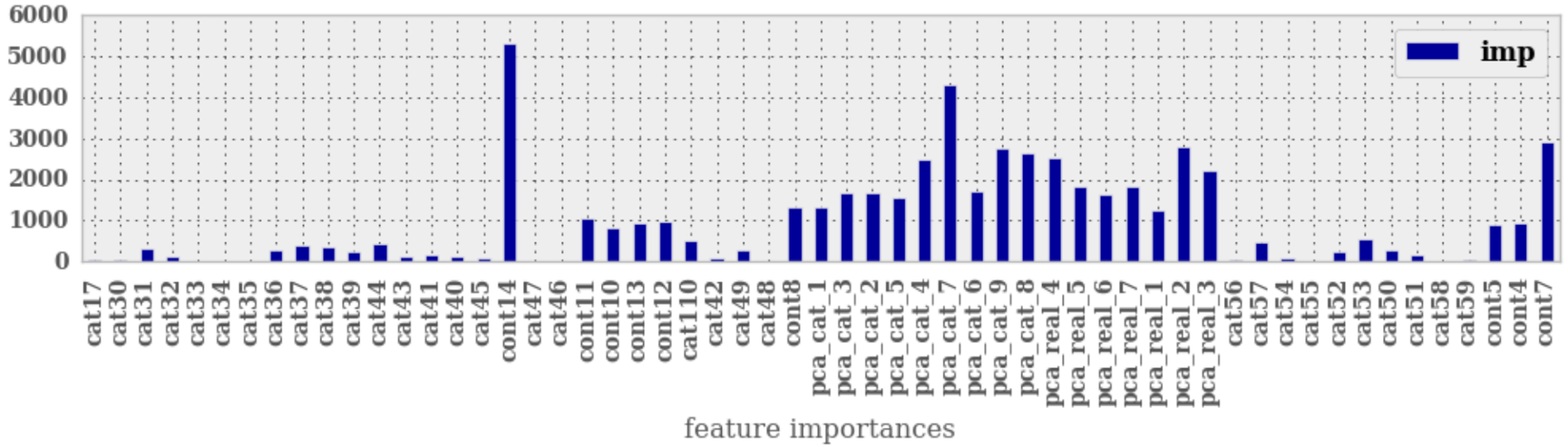
Правило столбцовых диаграмм

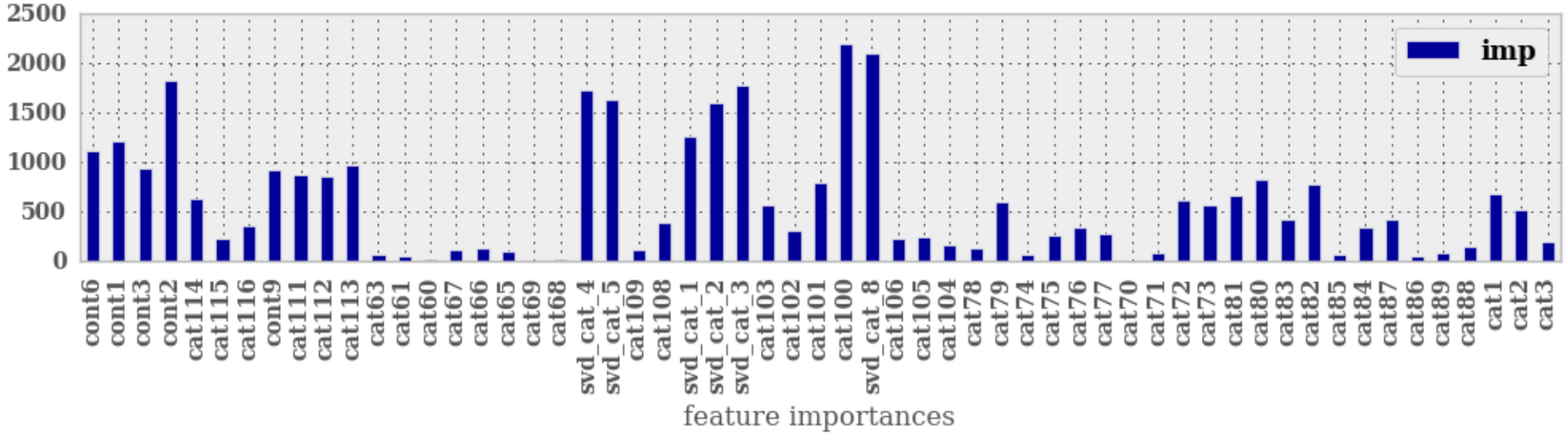
Правило:

- **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»

В начале решения задачи: 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)))
```

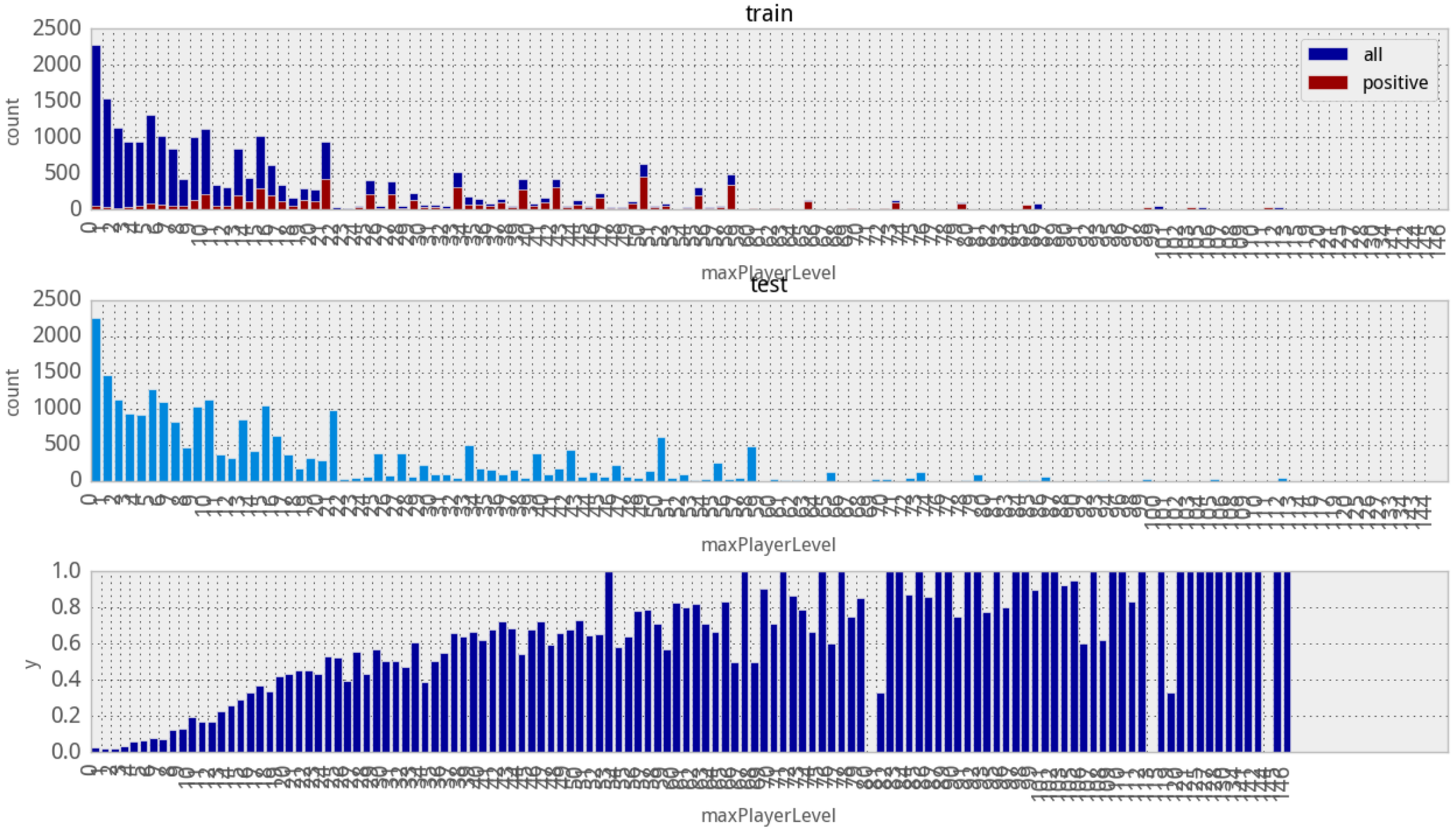
```
                    Класс          4       2.20       0.97 [1 2 3 4]
                    Номер       8404    7442.45     269.63 [5001 5002 ...]
                  Вес, т         124      38.27       7.30 [ 41.1  44.4  ...]
                   Начало       8404                  time
          Количество,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

Что надо сразу выяснить про признак

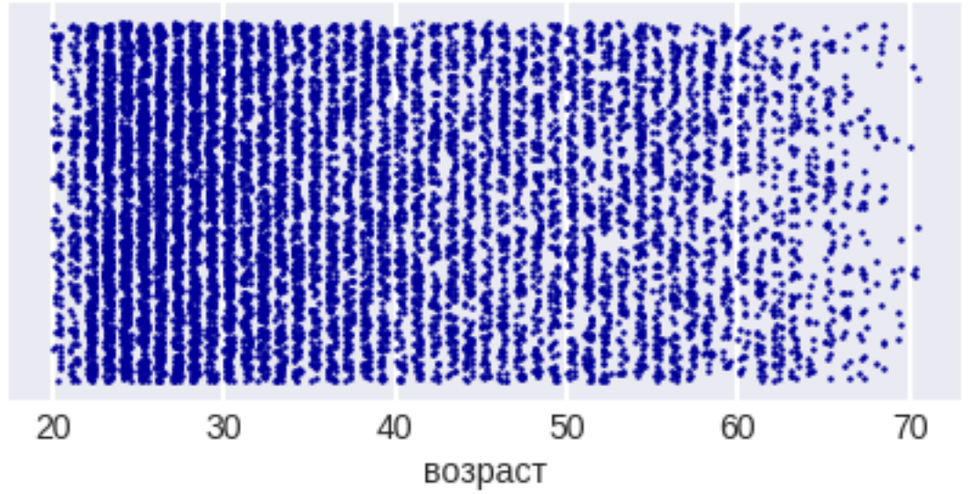
- распределение значений признака
- распределение обучение / тест
- распределение целевой переменной (ex: класс 0 / 1)

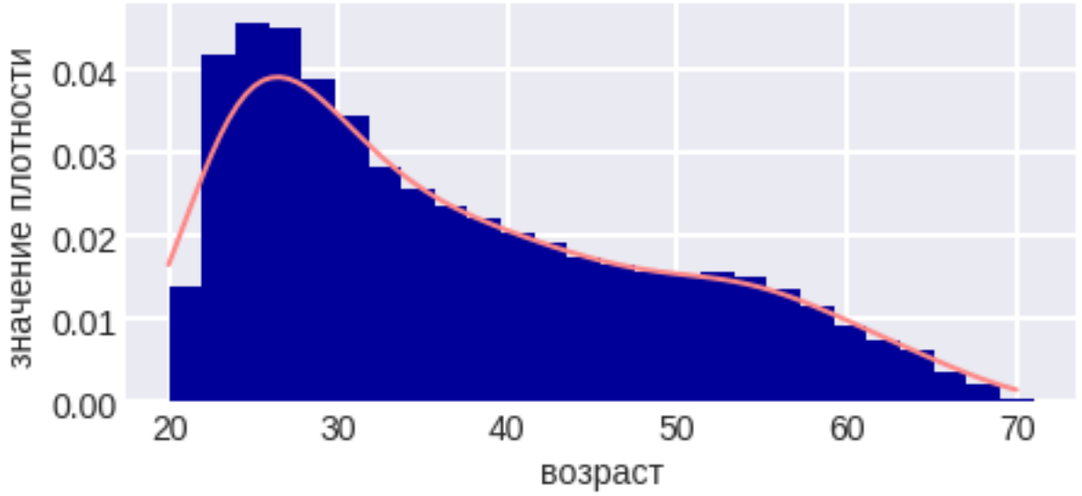
- такие же вопросы для пропусков, выбросов

Что надо сразу выяснить про признак



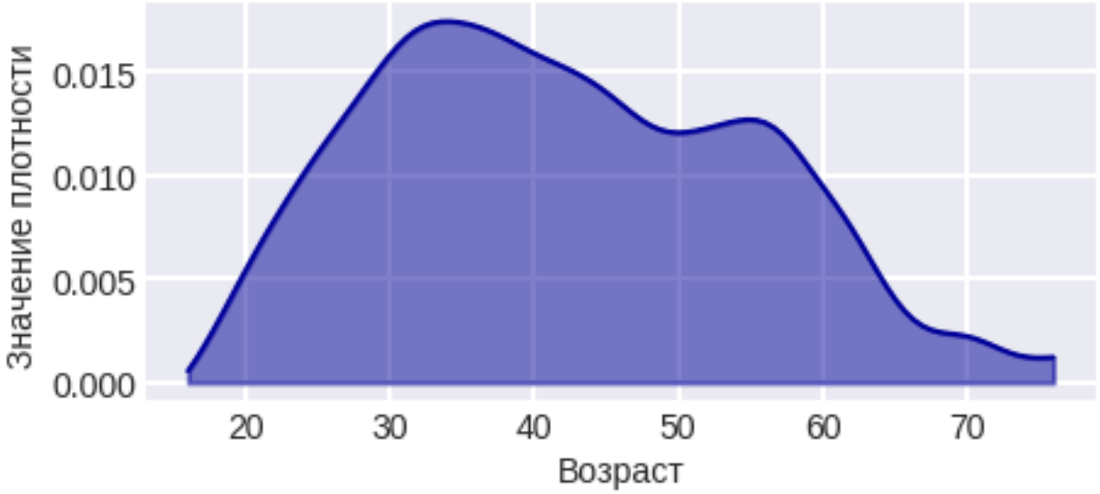
Визуализация отдельных признаков

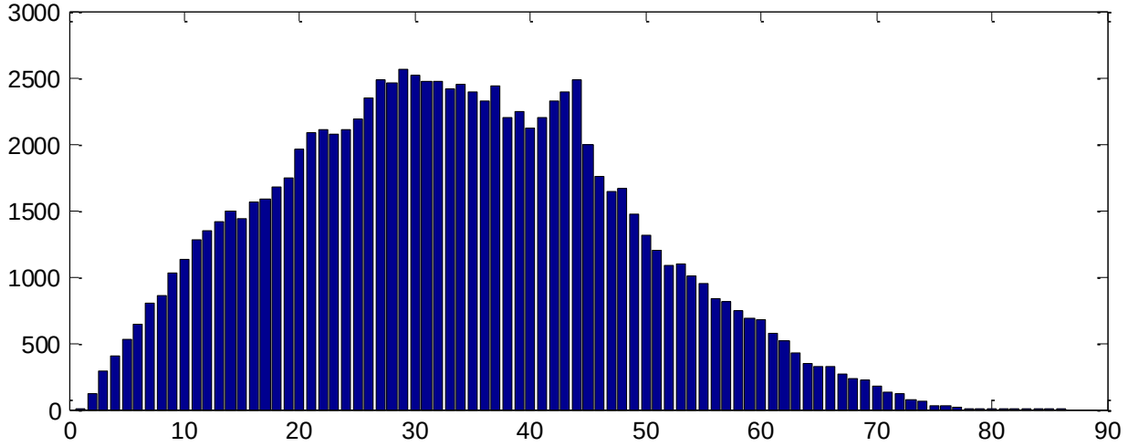




Гистограммы предпочтительнее плотностей

Задачи «М-магазин» / «ТКС»



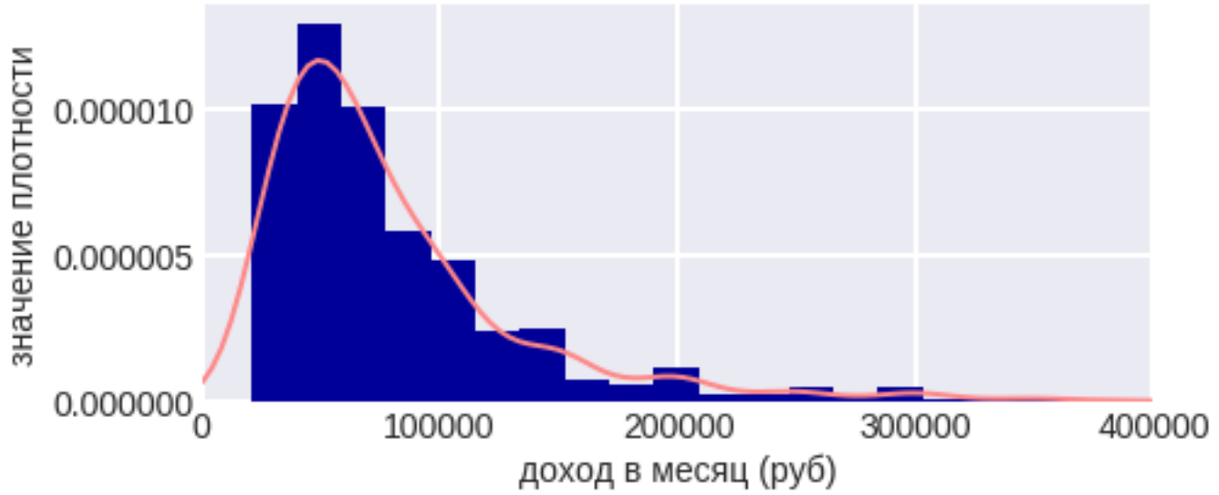


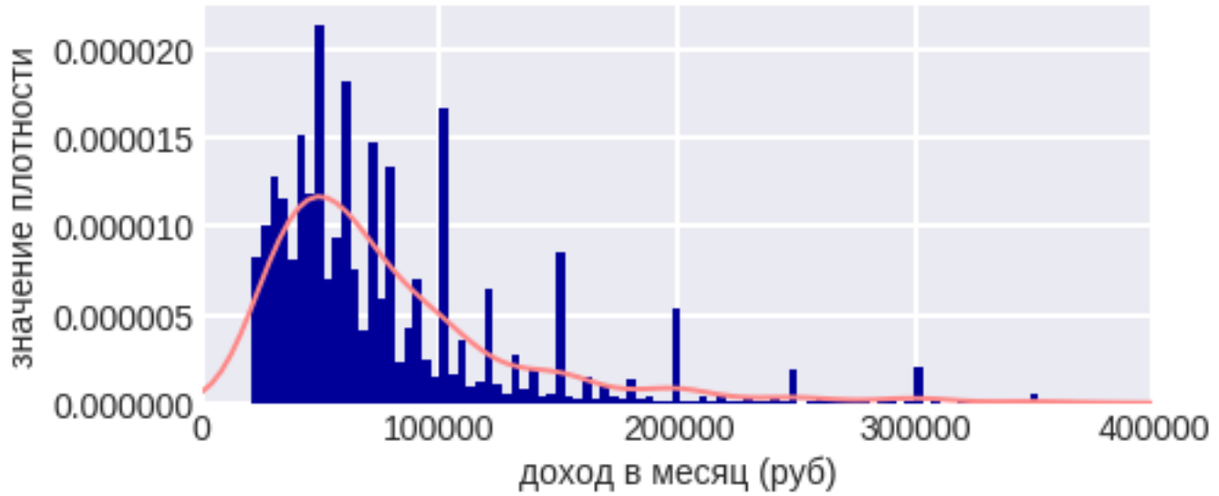
Распределение возраста покупателей

Так обычно выглядит распределение!

Почему два горба? / выступ?

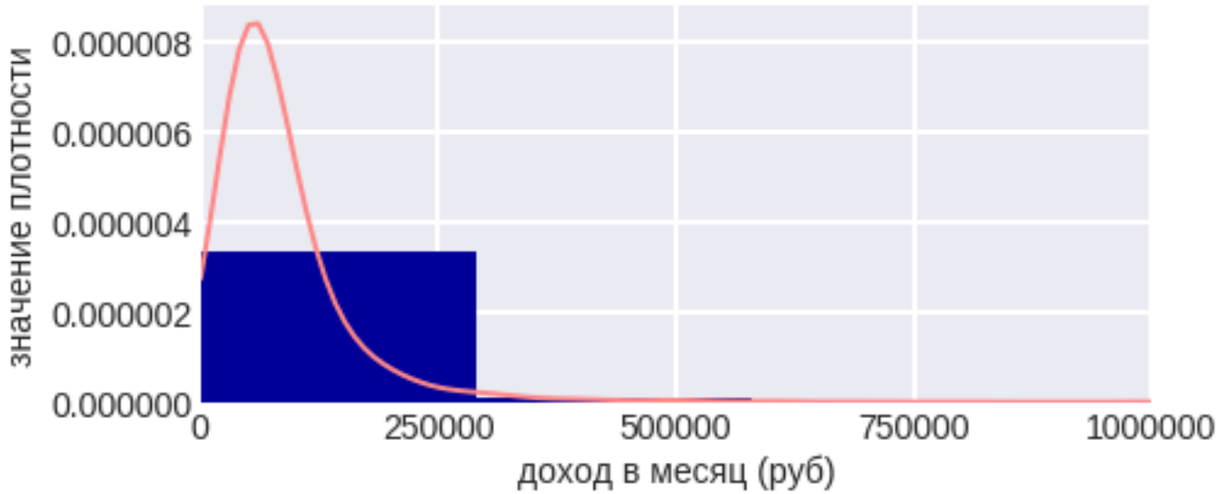
Проблемы визуализаторов – параметры по умолчанию





увеличили число бинов

Проблемы визуализаторов – выбросы



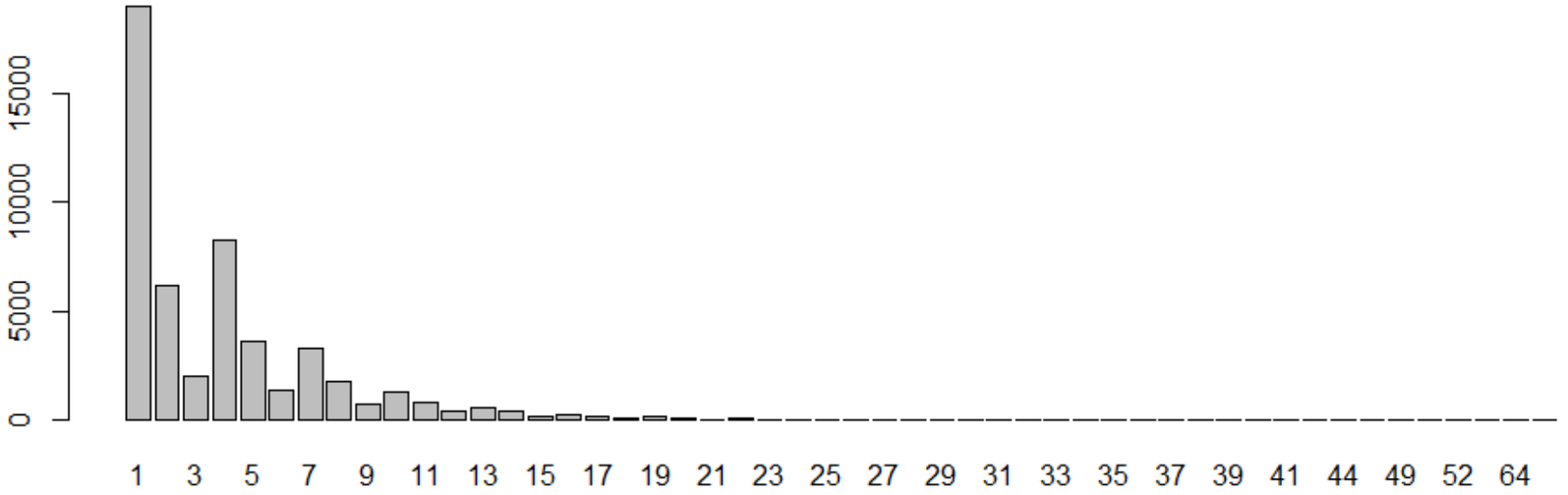
Что будет если не устранять выбросы…

```
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

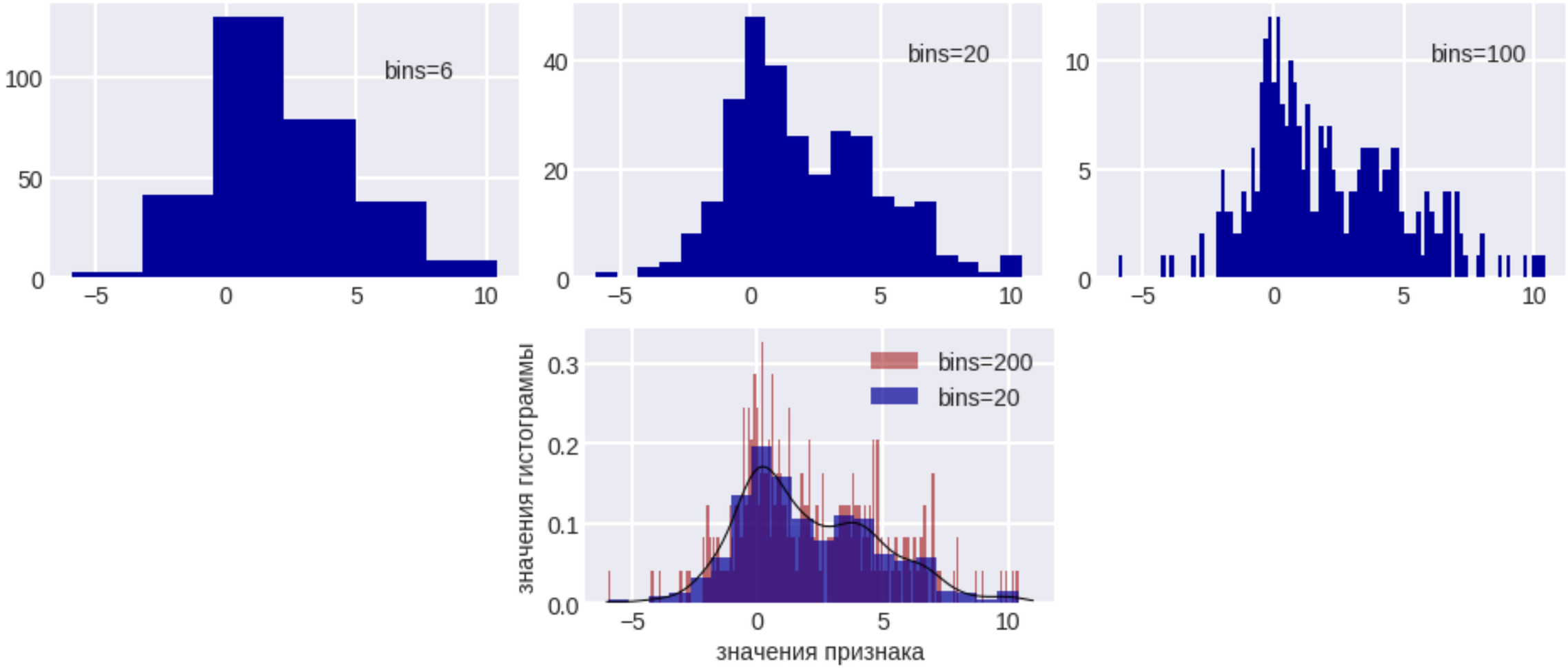
паттерны – «тройки»

Выбирать:

число бинов

ширина столбцов

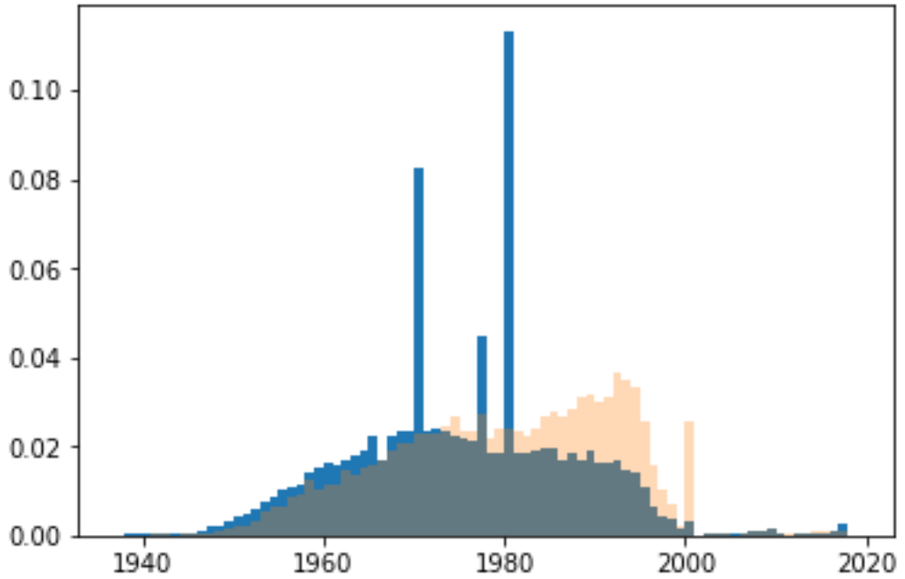
Построение гистограммы



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

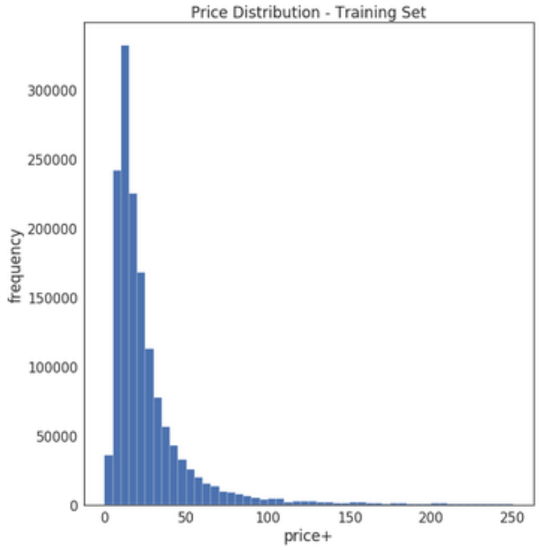
значения по умолчанию $\Rightarrow$ точная дата неизвестна

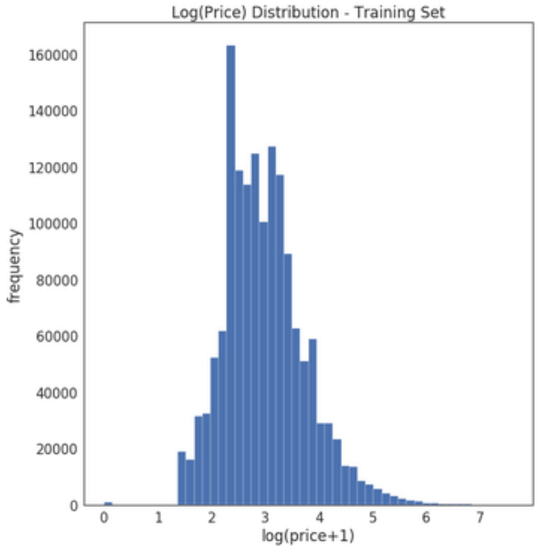
при этом пол «Ж» $\Rightarrow$ тоже неверно

Стоит ли доверять другой информации?

Выводы о признаках

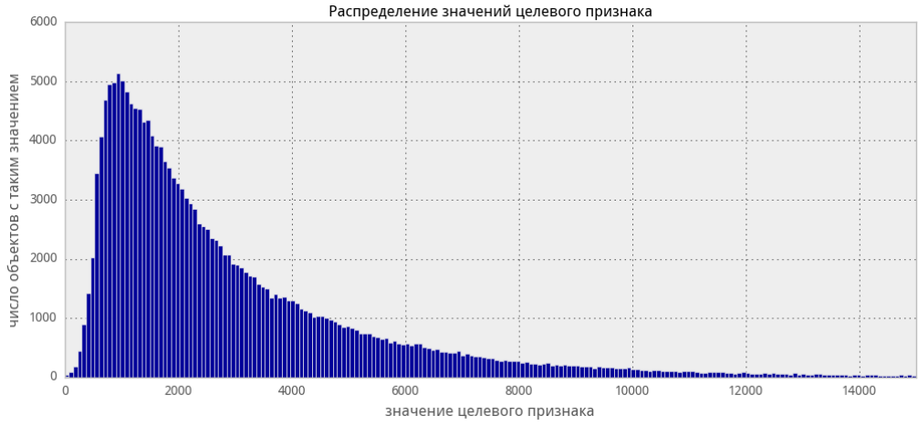
Использование визуализации для выбора трансформации





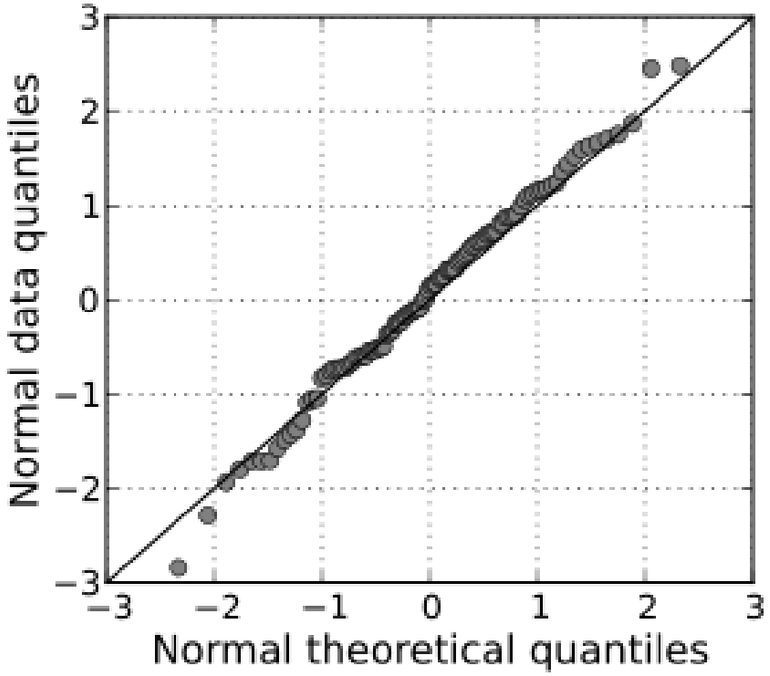
https://www.kaggle.com/thykhuely/mercari-interactive-eda-topic-modelling

«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!**

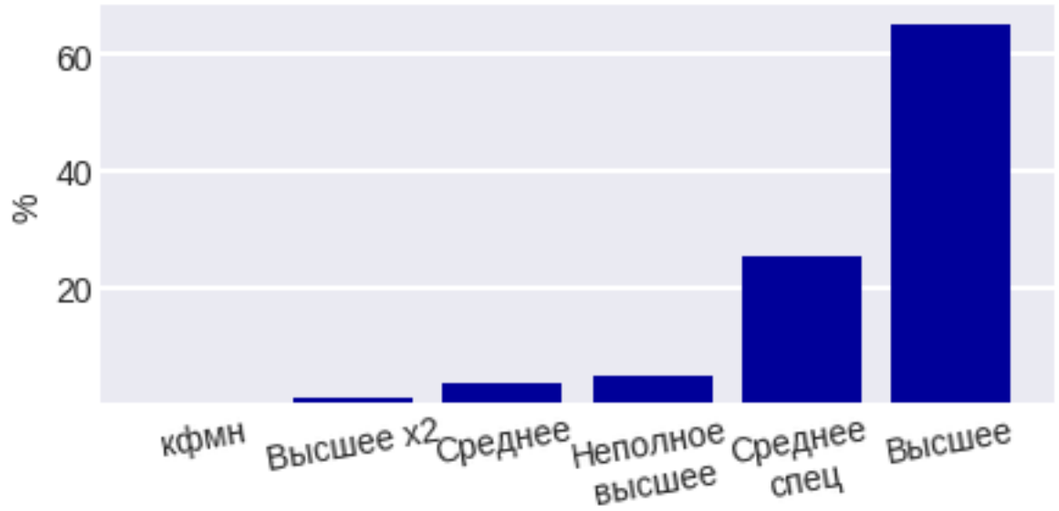
Зачем

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

Сравнение:

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

Визуализация категориальных признаков

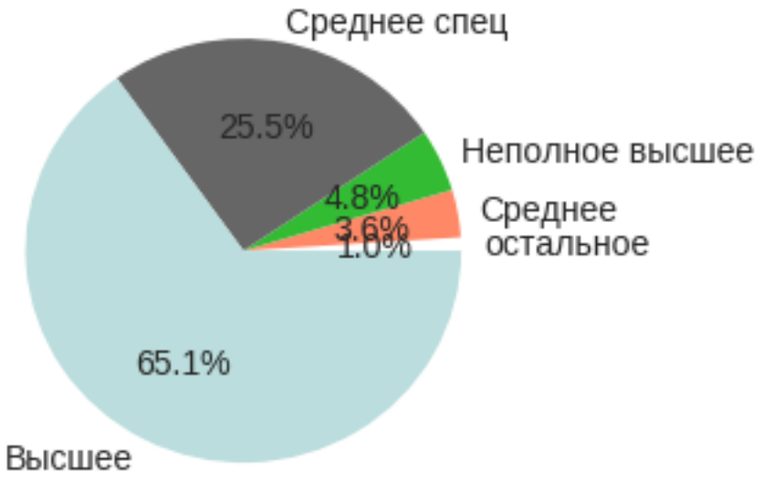


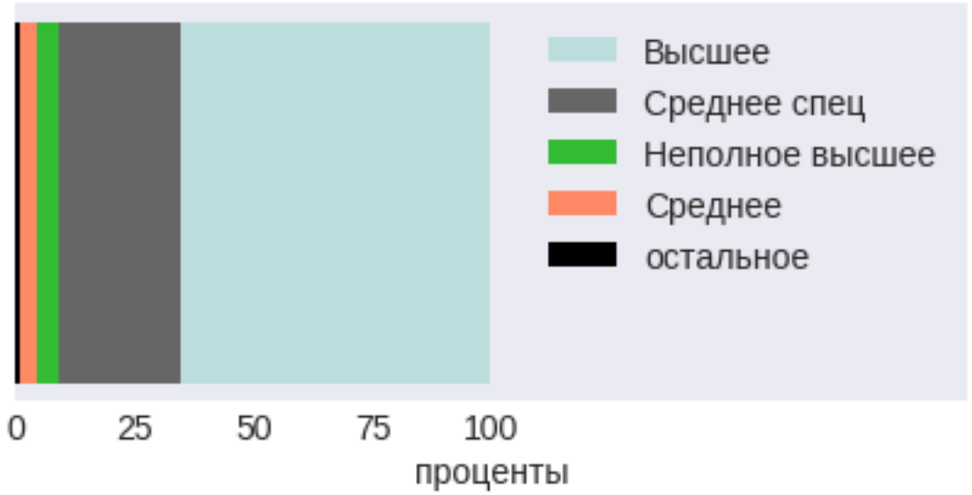
не видно мелкие категории

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

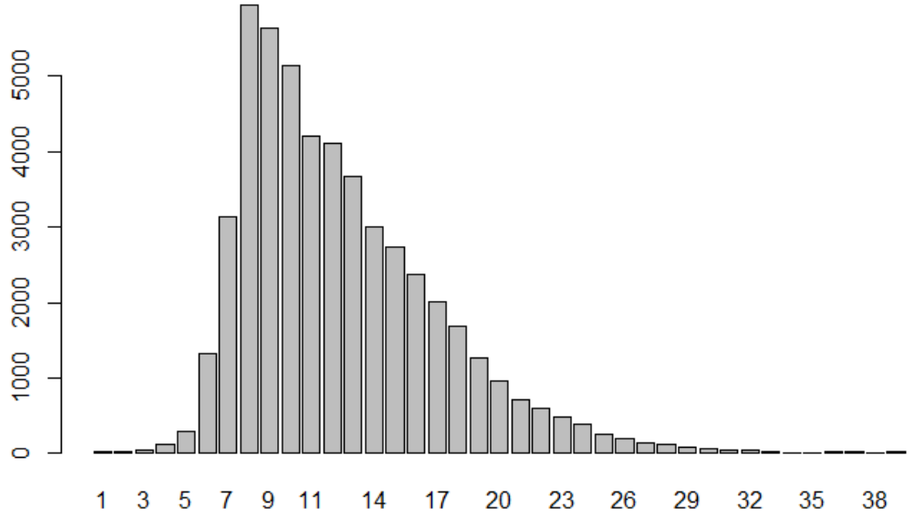
Когда информации для визуализации мало – таблицы!

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…

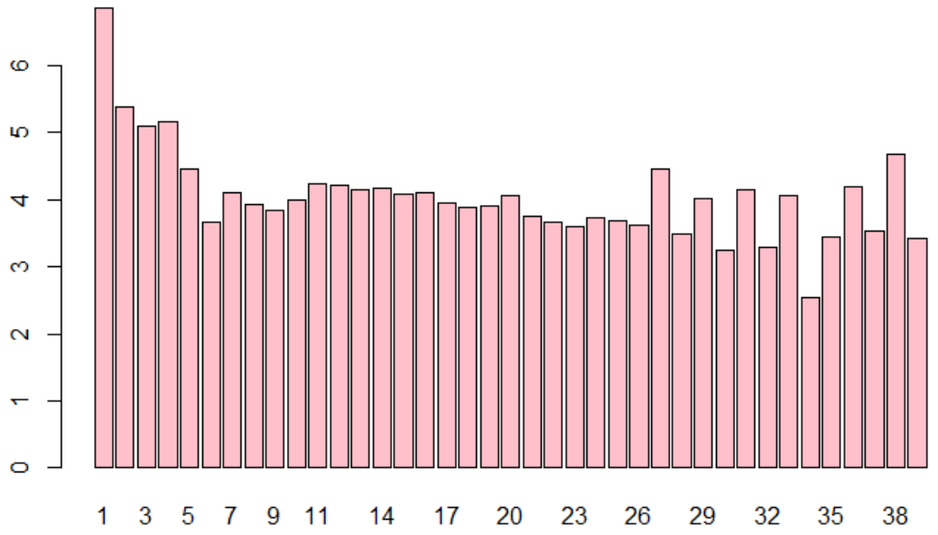
Распределения на признаках – природа признаков

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»

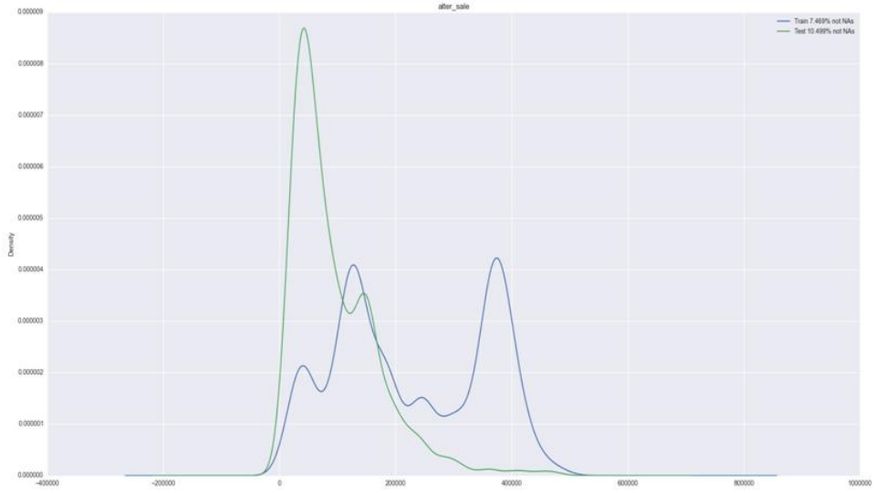
	mean	count
	cat101	
A	2454.139844	106721
B	1292.020000	3
C	2778.283638	16971
D	2812.990306	17171
E	4458.574286	7
F	3560.151861	10139
G	3450.680947	10944
H	1320.720000	1
I	4590.935254	6690
J	4603.863790	7259
K	3240.165000	2
L	5321.419556	3173
M	5540.292766	3669
N	2192.720000	1
O	6870.387172	2493
Q	7057.470264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

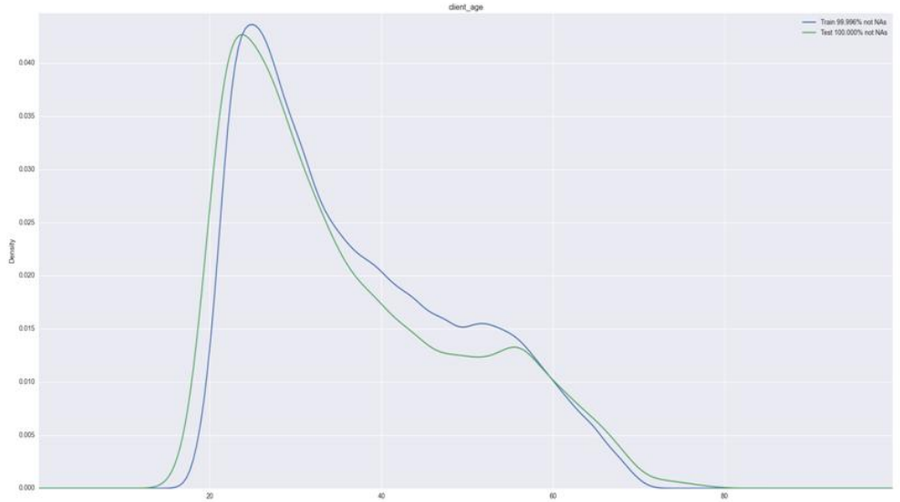
Как распределение меняется при переходе к контролю

смотреть как меняются распределения

обучение – контроль



Есть существенные изменения



Нет изменений

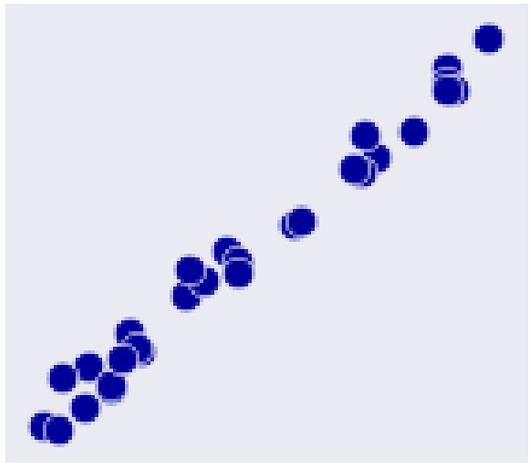
История про о-трэвел и волшебный признак.

Визуализация пары признаков

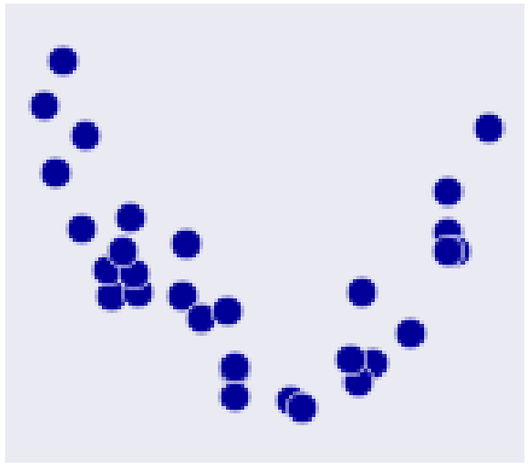
Самый распространённый способ –
диаграмма рассеивания («скатерплот»)

А что на диаграмме рассеивания 2х признаков можно увидеть?

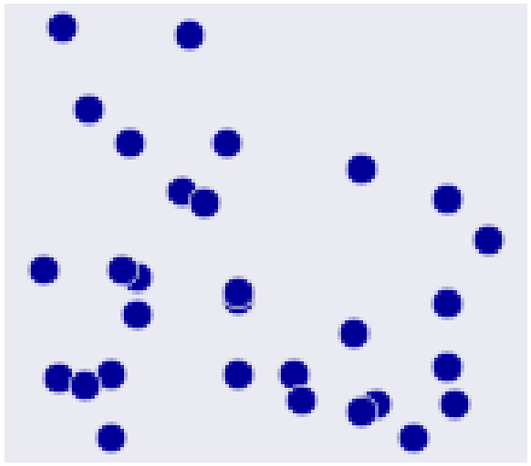
Что можно увидеть в данных («признак» – «признак»)



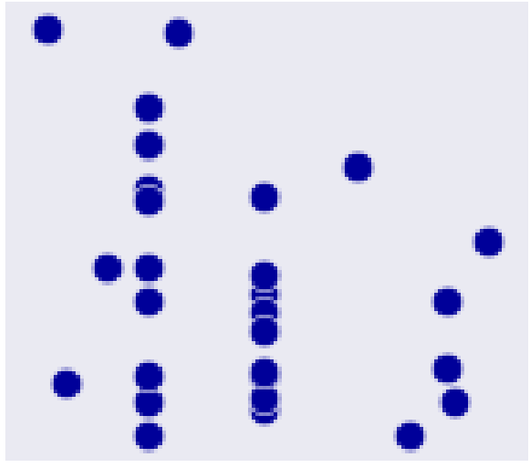
корреляция



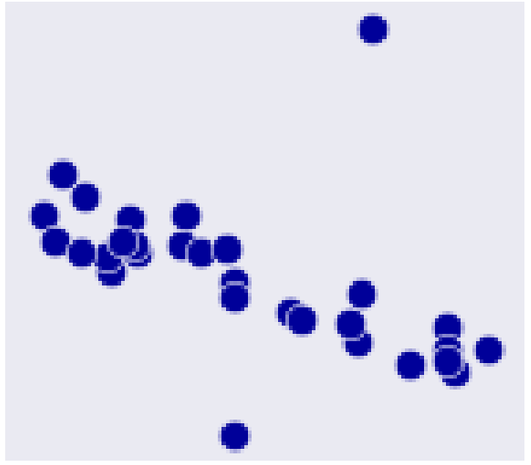
зависимость



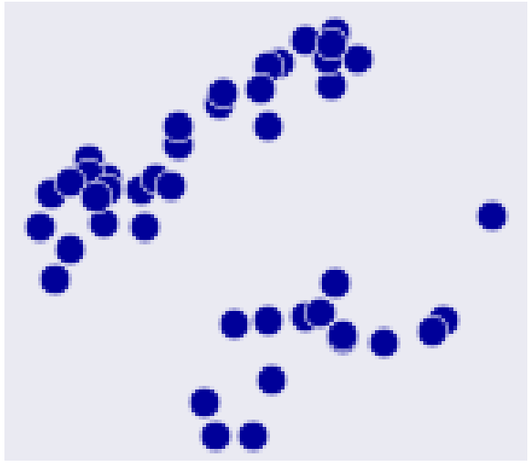
независимость



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

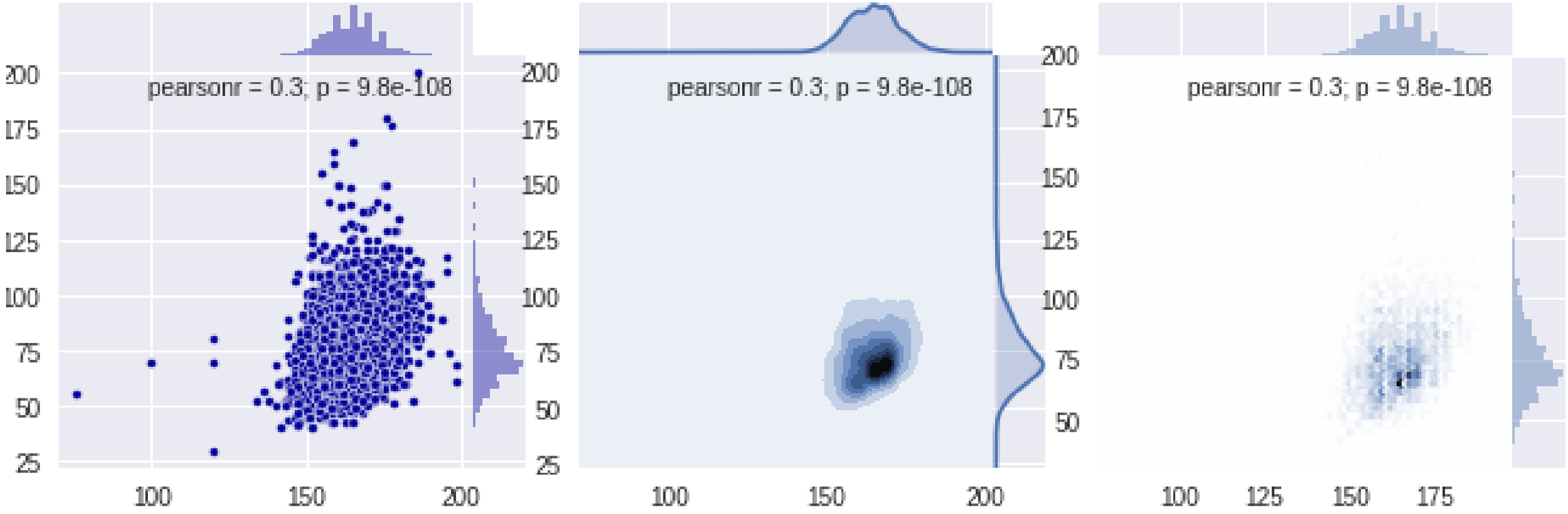
при правильном масштабе

кластеры

при правильном масштабе

Диаграмма рассеивания – лучший выбор

Задача о сердечно-сосудистых заболеваниях



признаки «рост-вес»

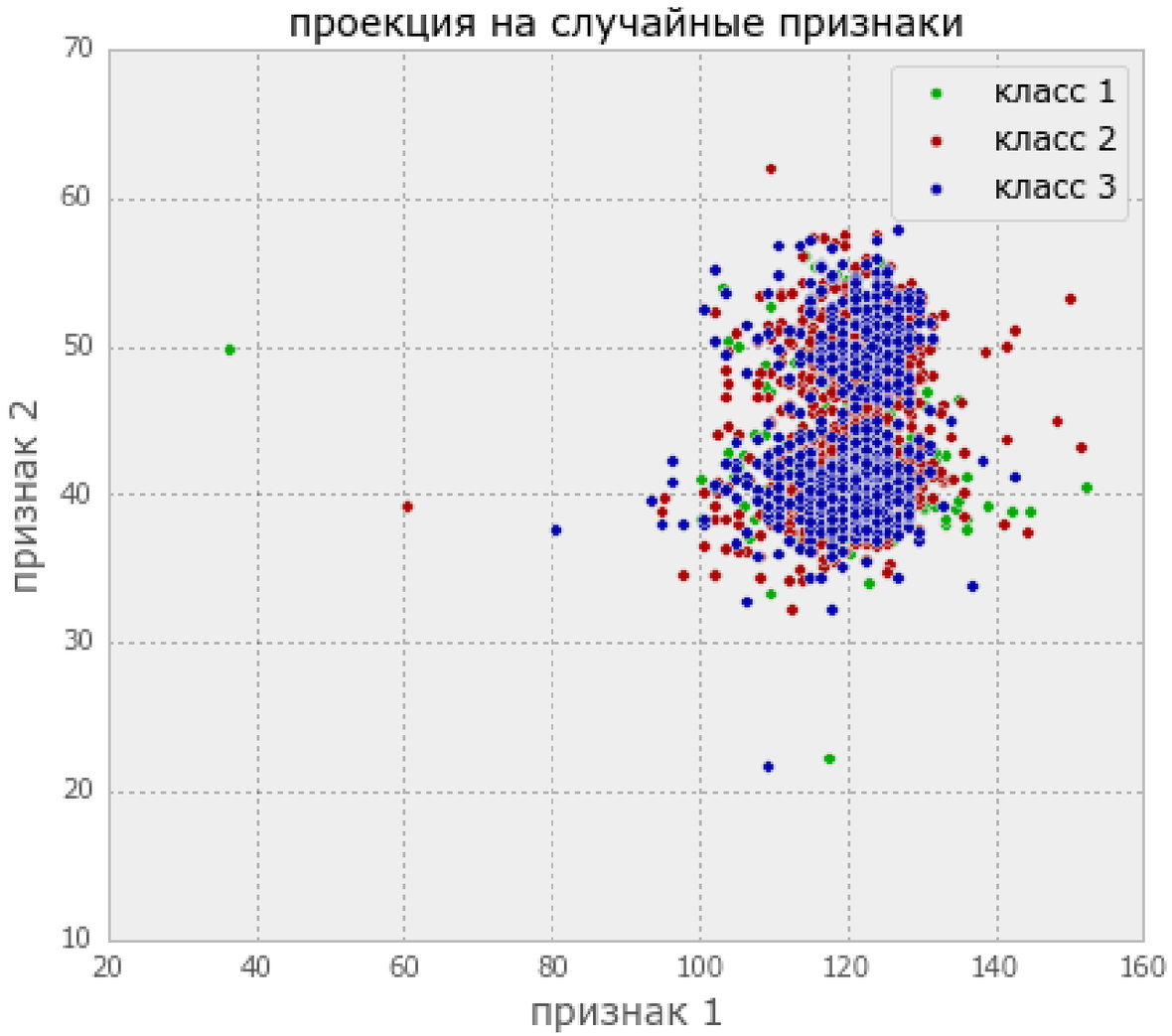
где видны выбросы?

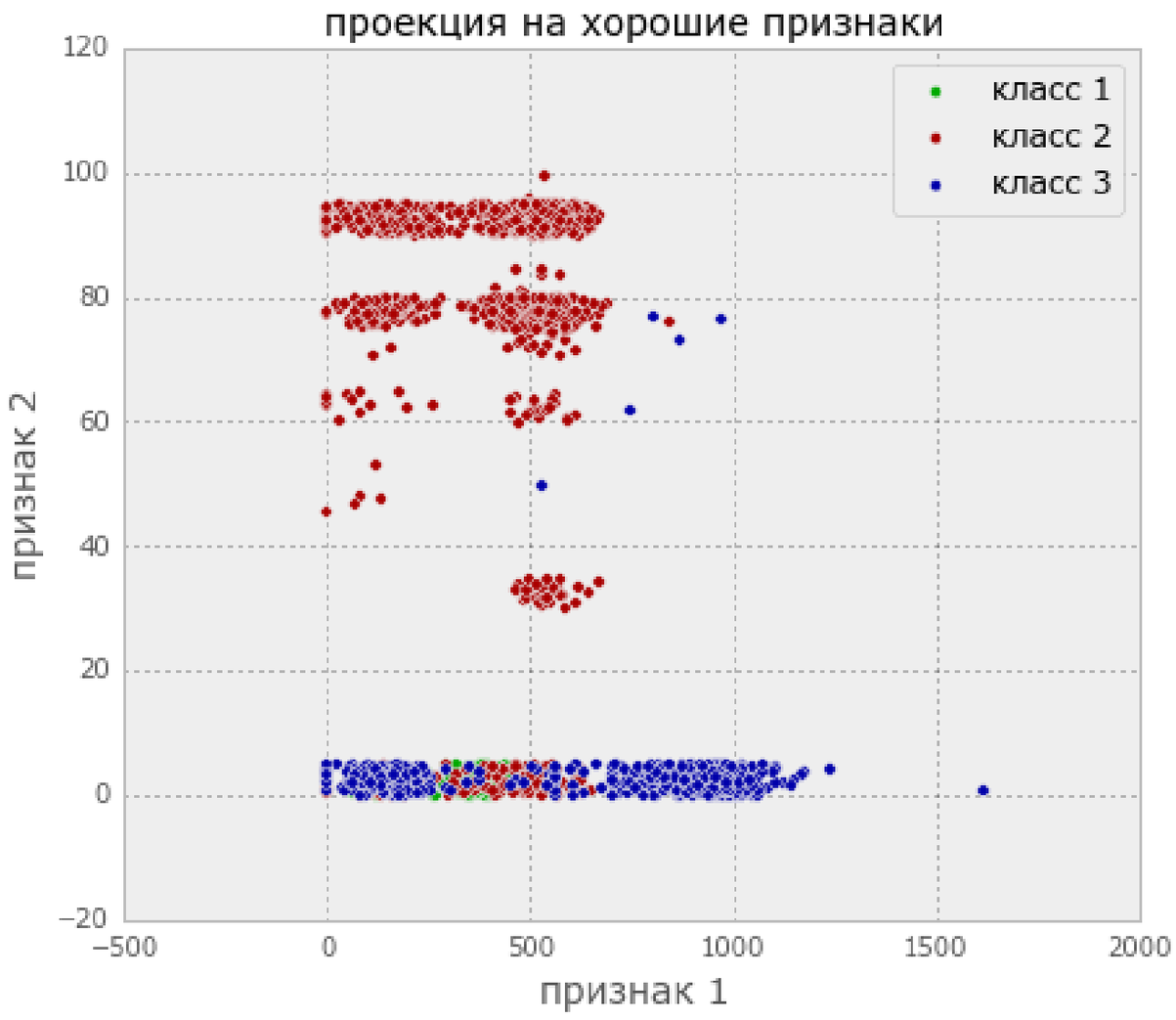
как сделать, чтобы и плотность анализировать?

Смотрим на пары признаков

- если есть время / признаков немного
- есть потенциально интересные сочетания

Смотрим на пары признаков

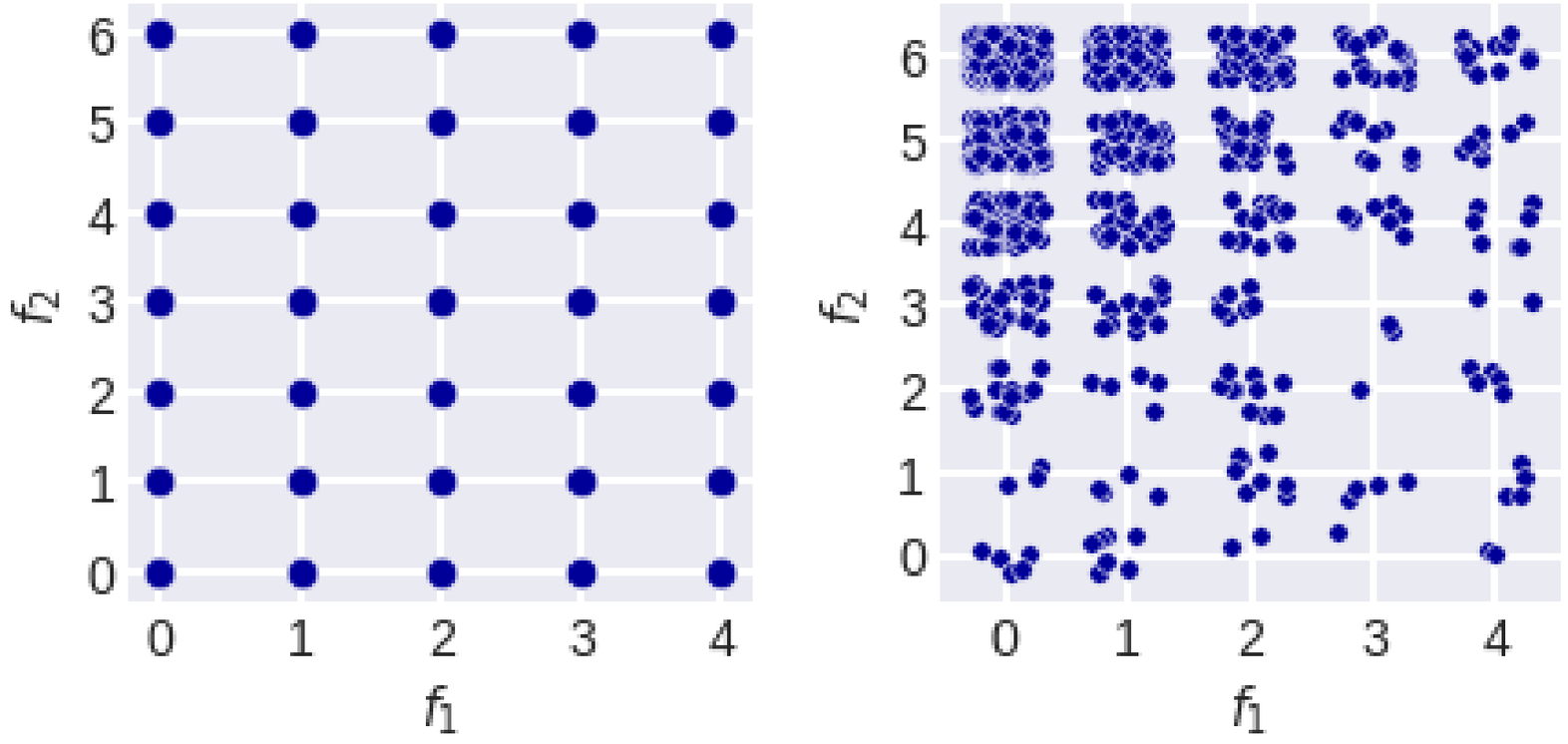




разница между случайными и хорошими признаками

Диаграммы рассеивания дискретных признаков

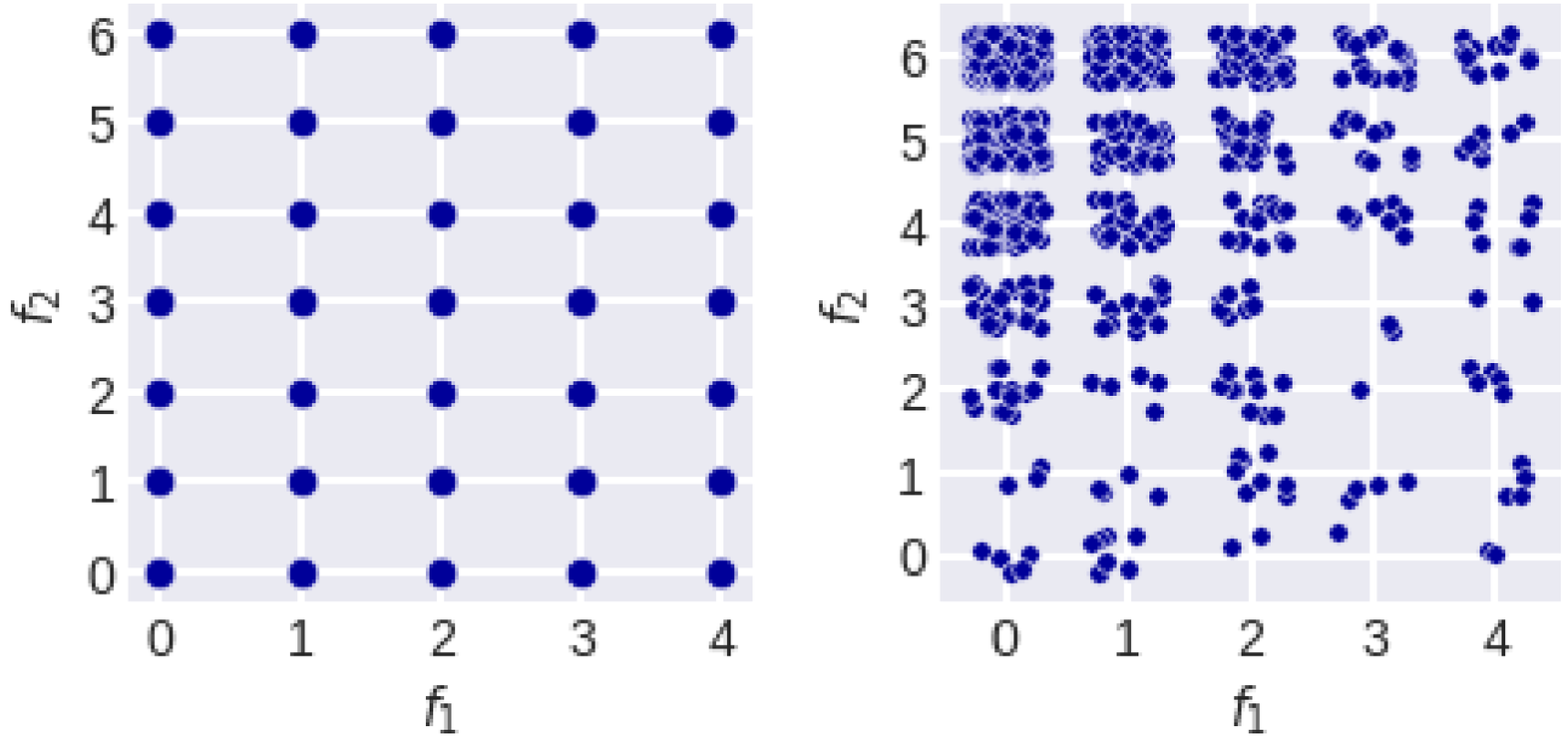
Зачем нужен Jitter



Что видно?

Диаграммы рассеивания дискретных признаков

Зачем нужен Jitter



Что видно?

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

Сводная таблица

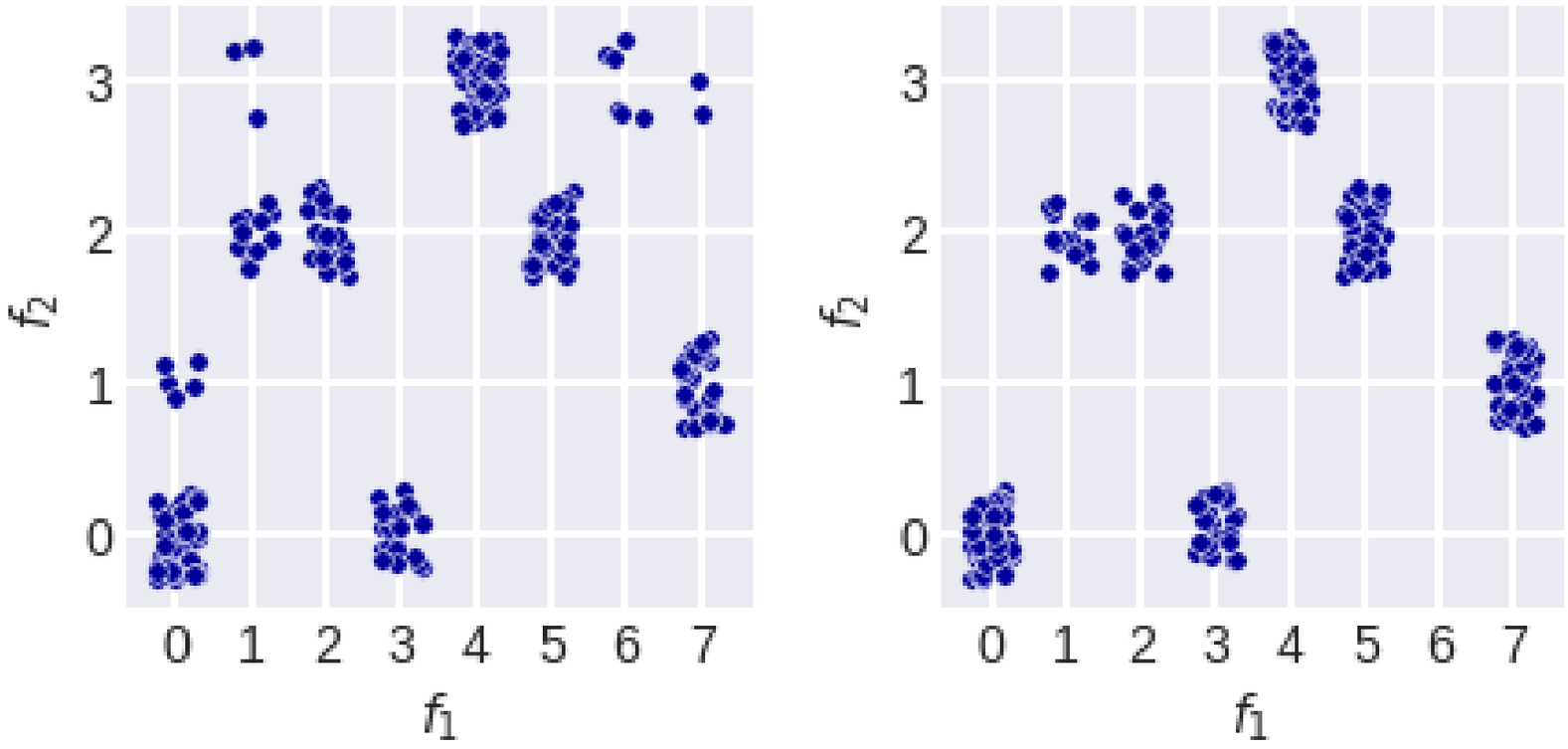
f_1 \ f_2	0	1	2	3	4	5	6
0	5	3	13	24	59	152	405
1	7	4	5	14	25	56	154
2	2	8	10	8	16	21	60
3	1	4	1	2	9	10	21
4	2	4	5	2	7	8	12

```
pd.crosstab(x1, x2)
```

Часто не нужно рисунков!

По таблице всё видно

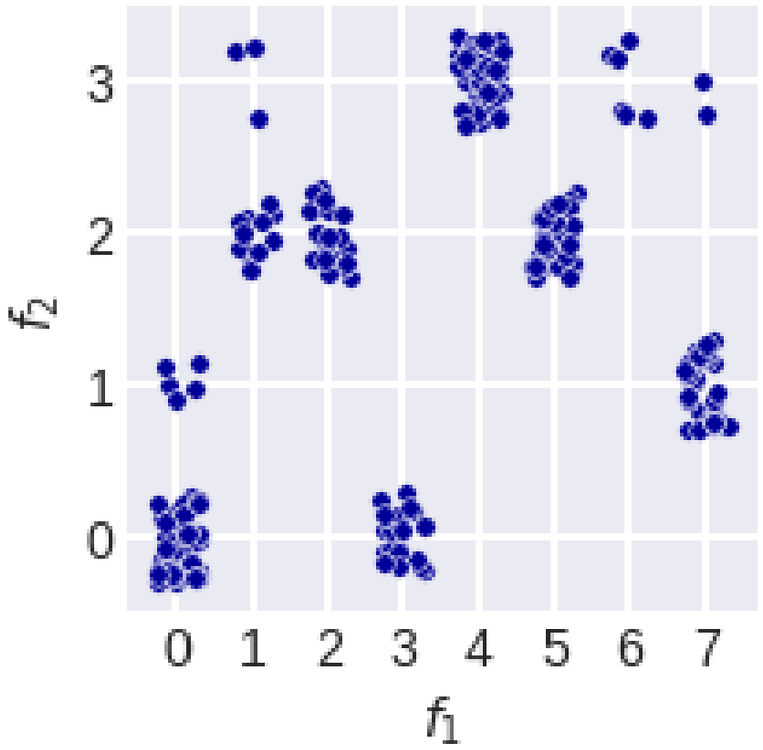
Диаграммы рассеивания дискретных признаков

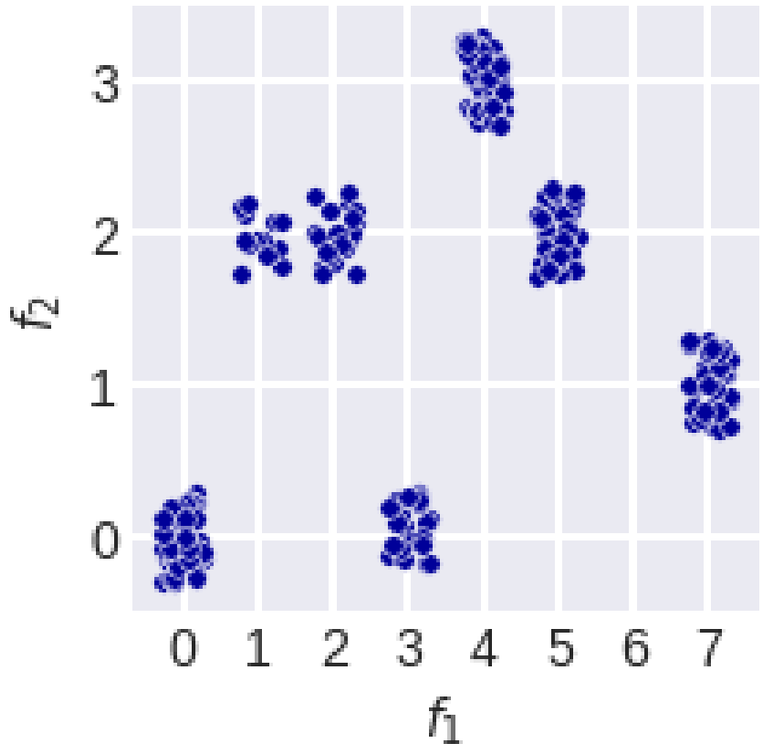


Справа – после удаления маленьких кластеров!

Что здесь видно?

Диаграммы рассеивания дискретных признаков



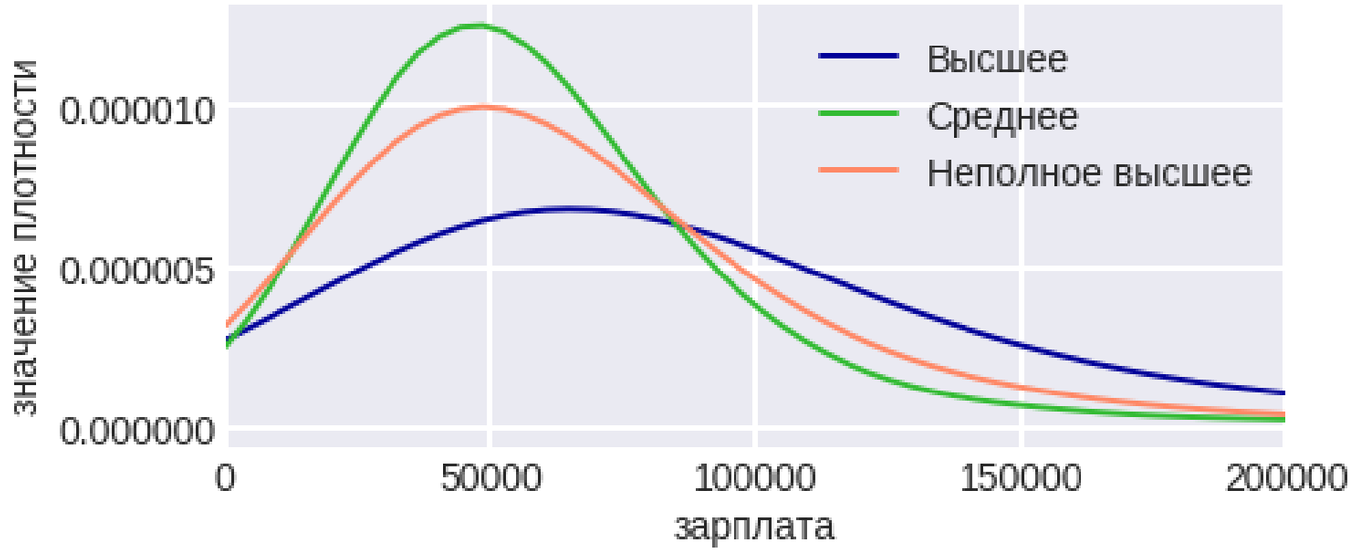


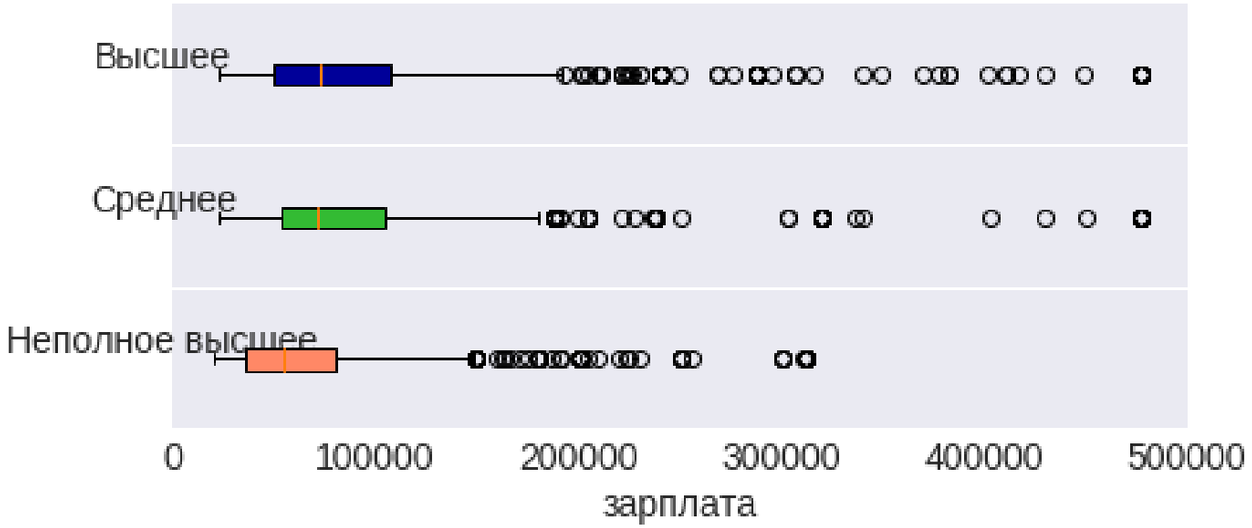
Один признак – уточнение другого!

Как это использовать?

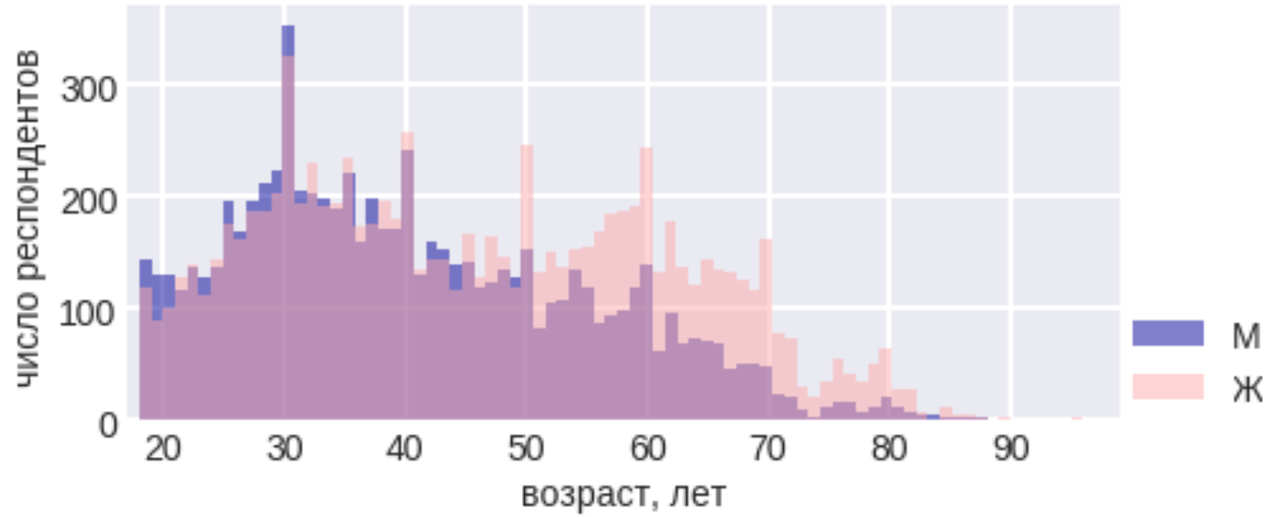
«Liberty»

Пара «вещественный признак – категориальный»

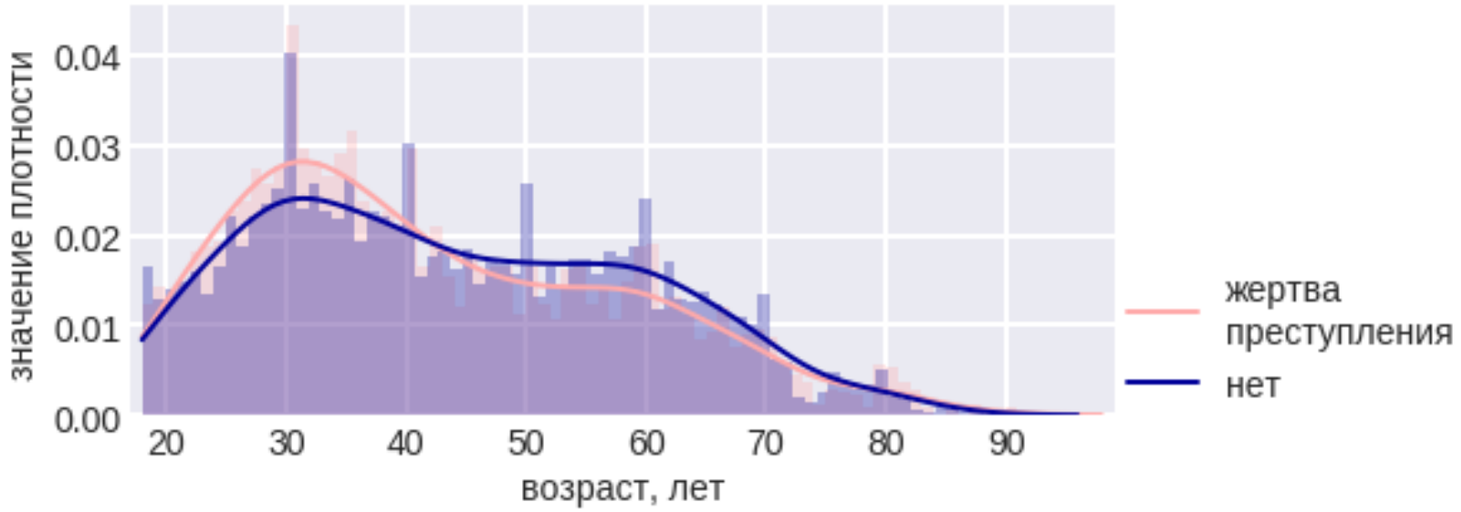




Пара «вещественный признак – категориальный»

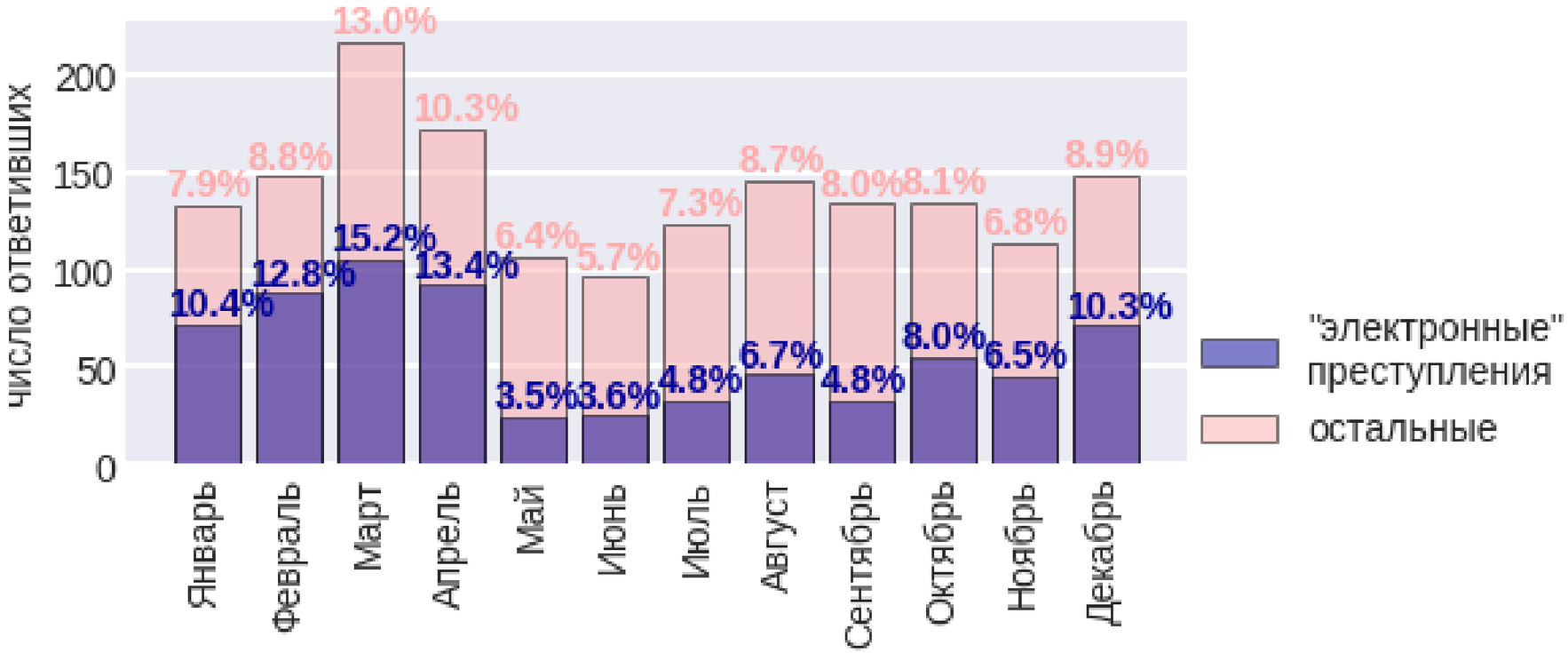


Распределение респондентов по возрасту и полу



Распределение возрастов жертв преступлений и остальных респондентов

Пара «бинарный признак – категориальный»

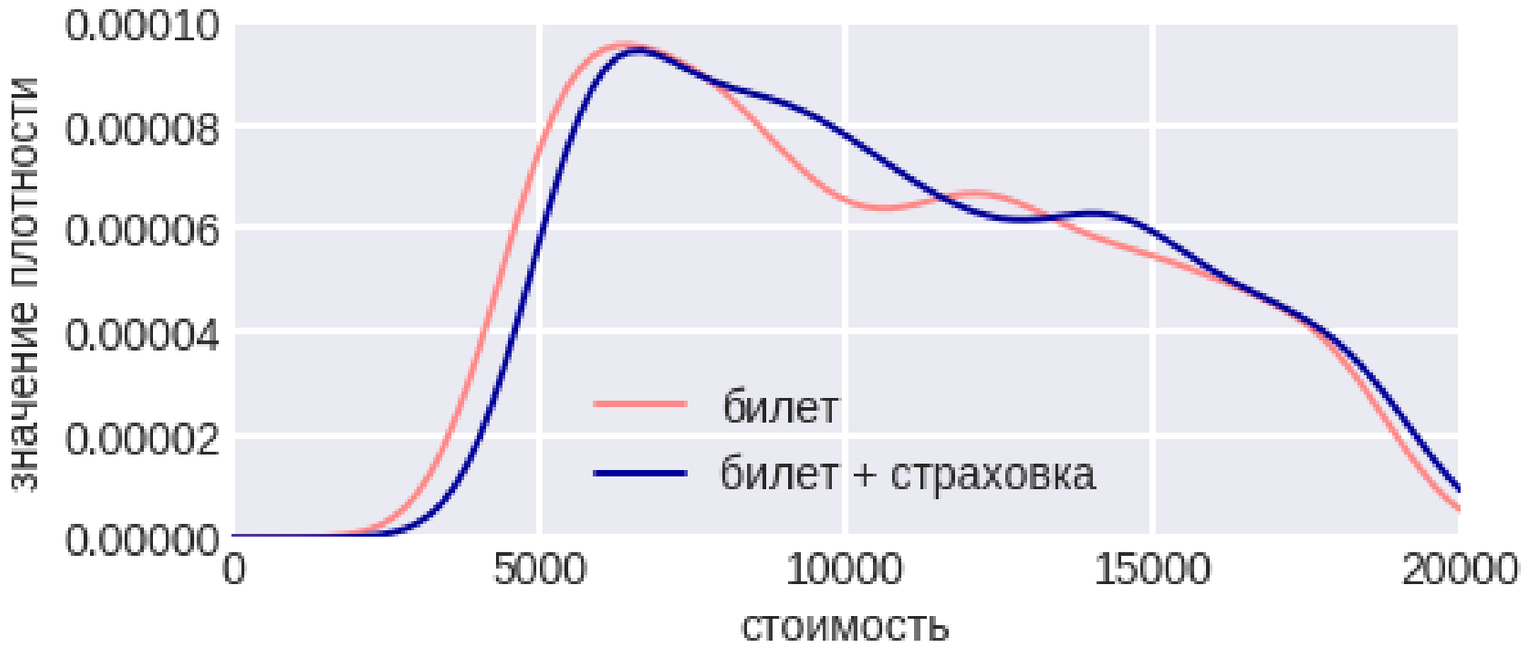


Здесь наоборот – по категориям средние значения бинарного

показан даже 3й признак – вид преступления

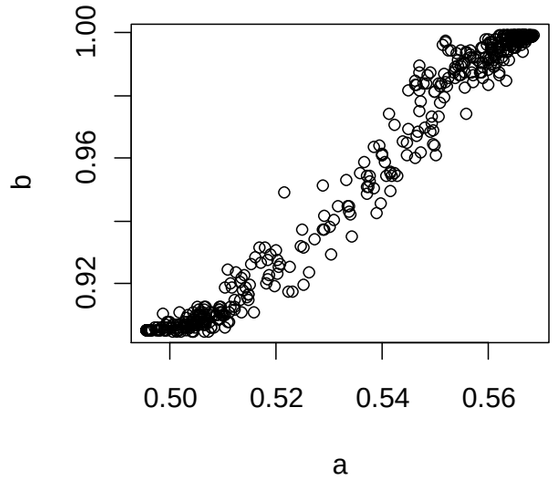
что видно из рисунка? какие выводы можно сделать?

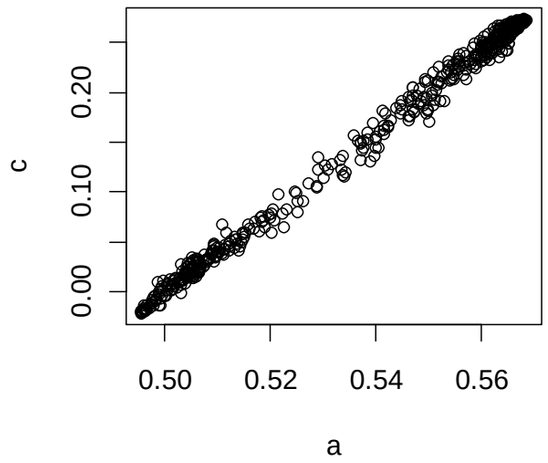
Задача «Ozon Travel»

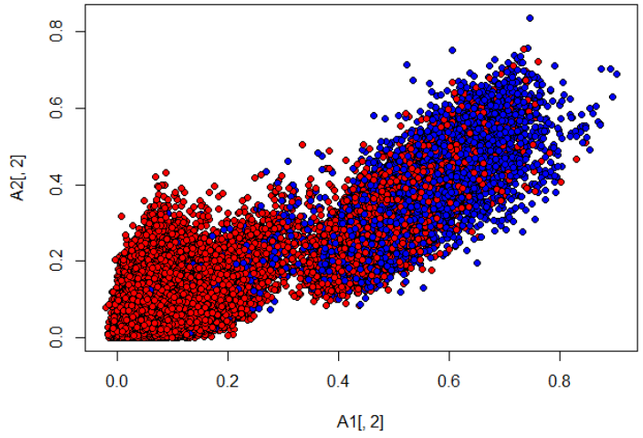


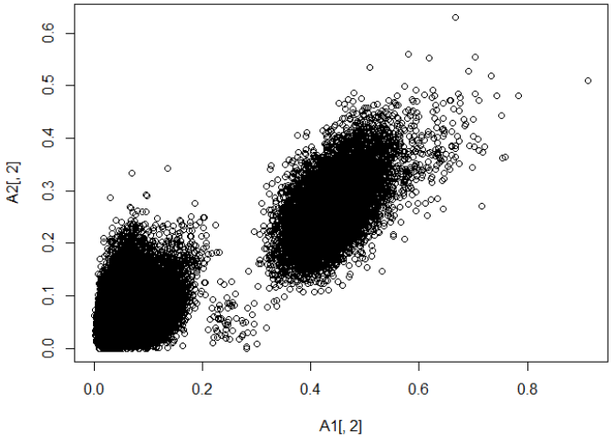
Всегда ставьте под сомнение свои выводы!

Визуализация ответов двух алгоритмов: как найти ошибку используя бенчмарк









Совет: создавайте бенчмарк!

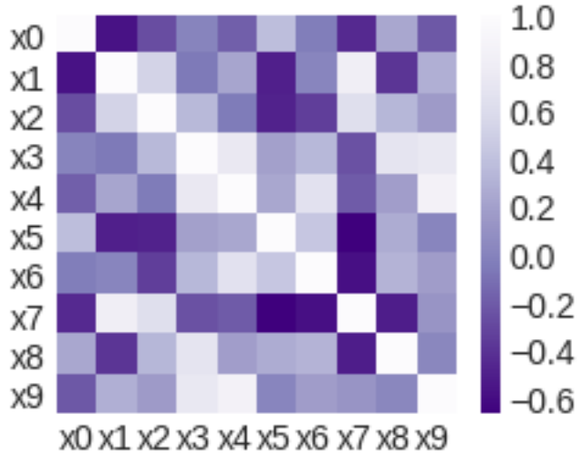
Корреляция между признаками

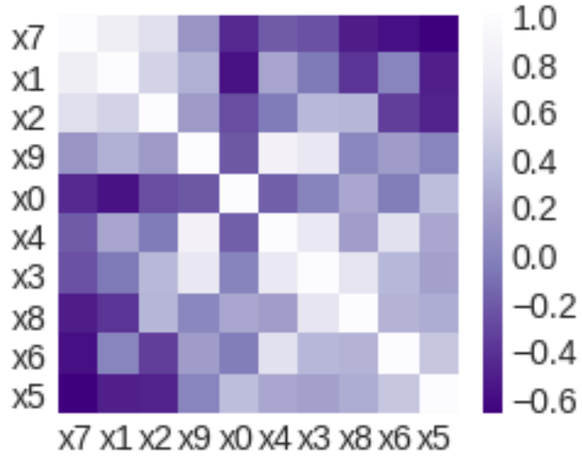
$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{m}(x_i - \operatorname{mean}(X))(y_i - \operatorname{mean}(X))}{m-1}$$

$$\operatorname{cov}(X,X) = \operatorname{var}(X) = \operatorname{std}^2(X)$$

$$\operatorname{cor}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\operatorname{std}(X)\operatorname{std}(Y)}$$

Корреляция между признаками






```
plt.imshow(df.corr(),
           interpolation='none')
plt.colorbar()
```

```
from scipy.sparse.linalg import svds
ii, _, _ = svds(cr, k=1)
ii = np.argsort(ii[:,0])
plt.imshow(cr.iloc[ii, ii],
           interpolation='none')
```

Такие матрицы сложно анализировать – требуется упорядочить

Корреляция между признаками

Корреляция – линейная зависимость…

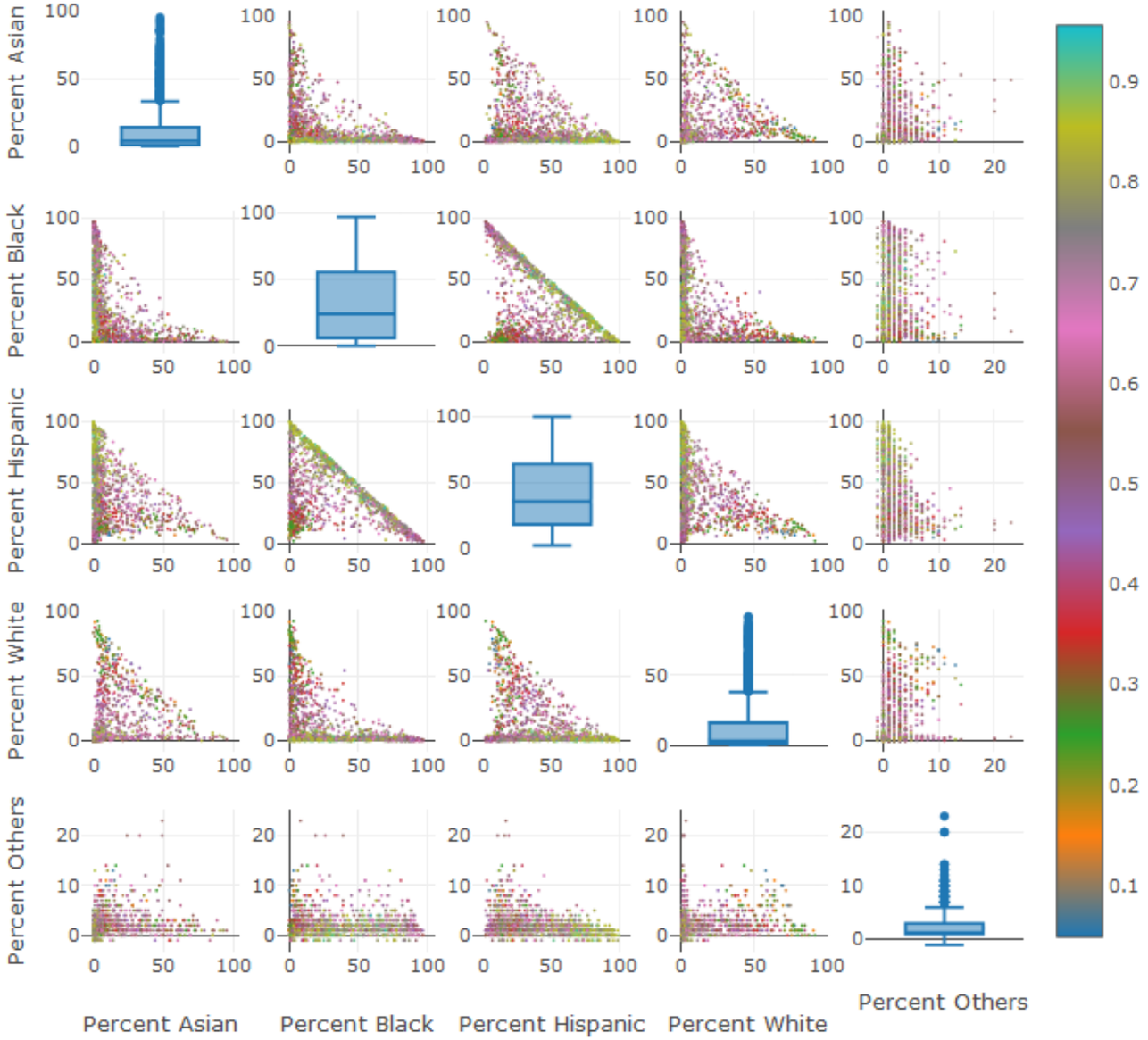
Можно

- **нелинейные (как?)**
- **характеристические векторы пропусков**
- **ранговые корреляции**

Как сгенерировать картинки с разными коэффициентами корреляции?

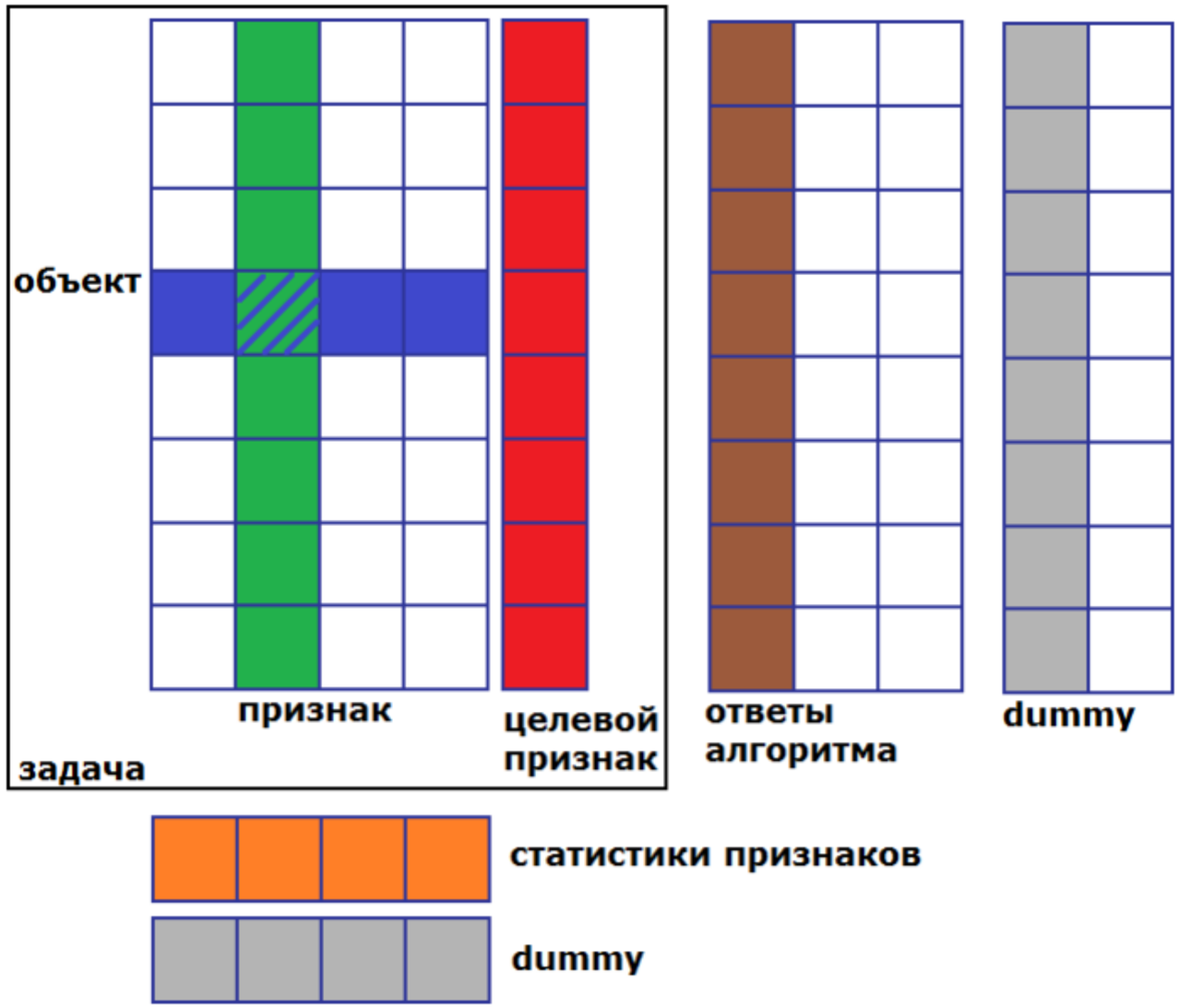
см. https://en.wikipedia.org/wiki/Correlation_and_dependence

Информация по всем парам – как правило, сильно перегружена



https://www.kaggle.com/thebrownviking20/passnyc-eda-and-unsupervised-learning

Что можно визуализировать



Что можно визуализировать

«Всё вертикальное»

- **признаки** (как исходно заданные, так и сгенерированные)
- **целевой признак**
- **ответы алгоритмов** (train – OOB-ответы, test – ответы)
- **служебные признаки** («нелогичные»: номер строки, случайный столбец, категория данных: обучение, валидация или тест и т.п.)

«Всё горизонтальное» (реже)

- **объекты или измерения**
- **статистики признаков**
- **служебная информация** (номера признаков, их категории и т.п.)

Итог

Гистограммы очень хороши

- быстро оценить форму распределения
- придумать деформацию

но надо настраивать вручную (впрочем, любую визуализацию)

Смотреть по признакам

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

Приёмы:

деформация признака (чаще логарифмирование)

масштабирование

- **не используйте сложных средств визуализации**

Досконально понимать, как происходит сама визуализация!

Итог

- **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)

визуализируйте всё вертикальное и горизонтальное

ищите объяснение всему, что видите на картинке

+ придумывайте, как это использовать для ML

понимайте достоинства и недостатки (что скрывает) конкретного типа визуализации!

данные важнее картинки!

храните данные

визуализация не должна быть лучше данных...