курс «Прикладные задачи анализа данных»

Искусство визуализации

Часть 3. Многомерный анализ

Александр Дьяконов

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?»

Многомерный анализ (multivariate analysis)

– анализируем две или более переменных

визуализация пары признаков (для разных типов)

визуализация ответов алгоритма

корреляции между признаками

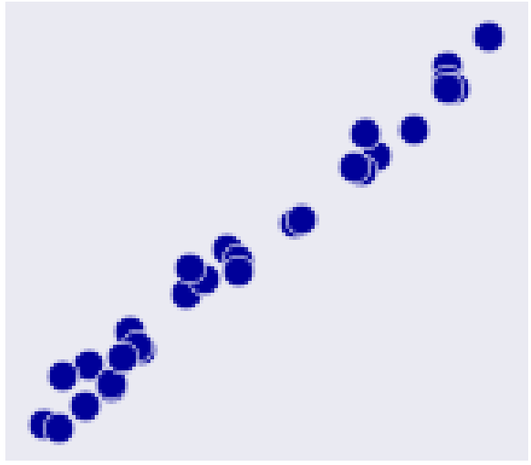
что можно визуализировать в табличных данных

Визуализация пары признаков

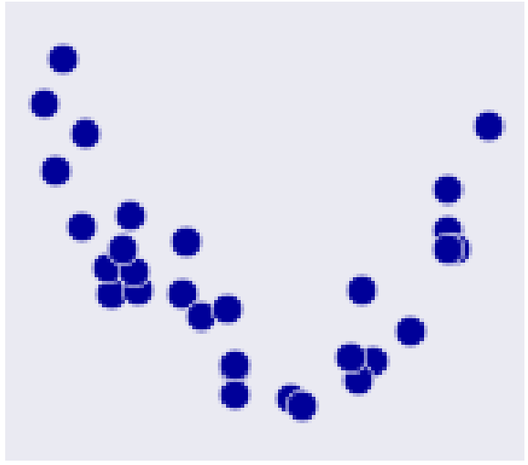
Самый распространённый способ –
диаграмма рассеивания («скатерплот»)

А что на диаграмме рассеивания 2х признаков можно увидеть?

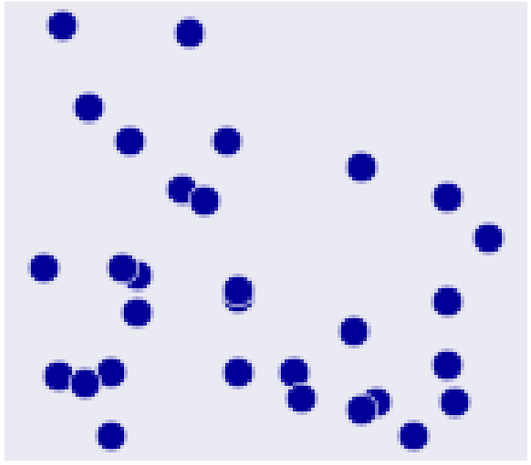
Что можно увидеть в данных («признак» – «признак»)



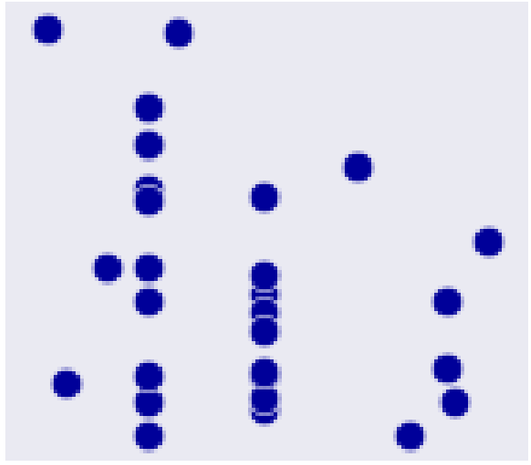
корреляция



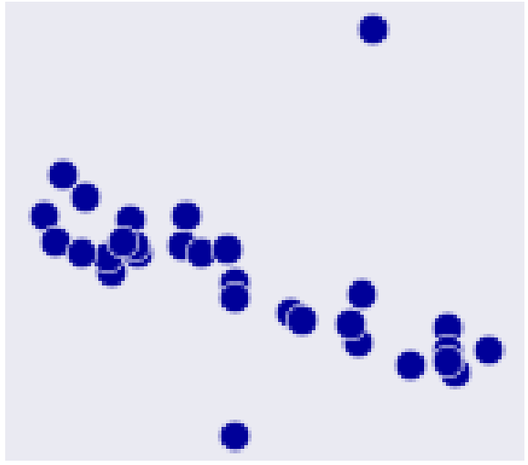
зависимость



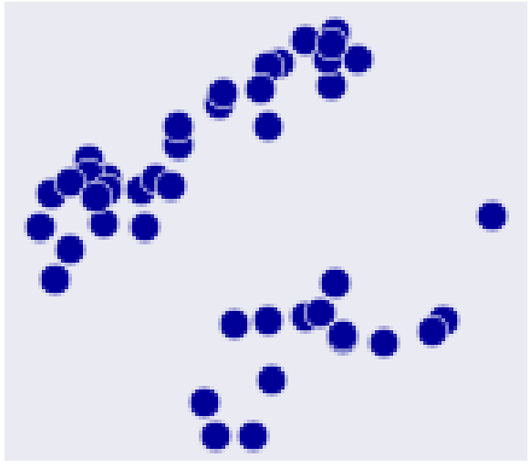
независимость



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

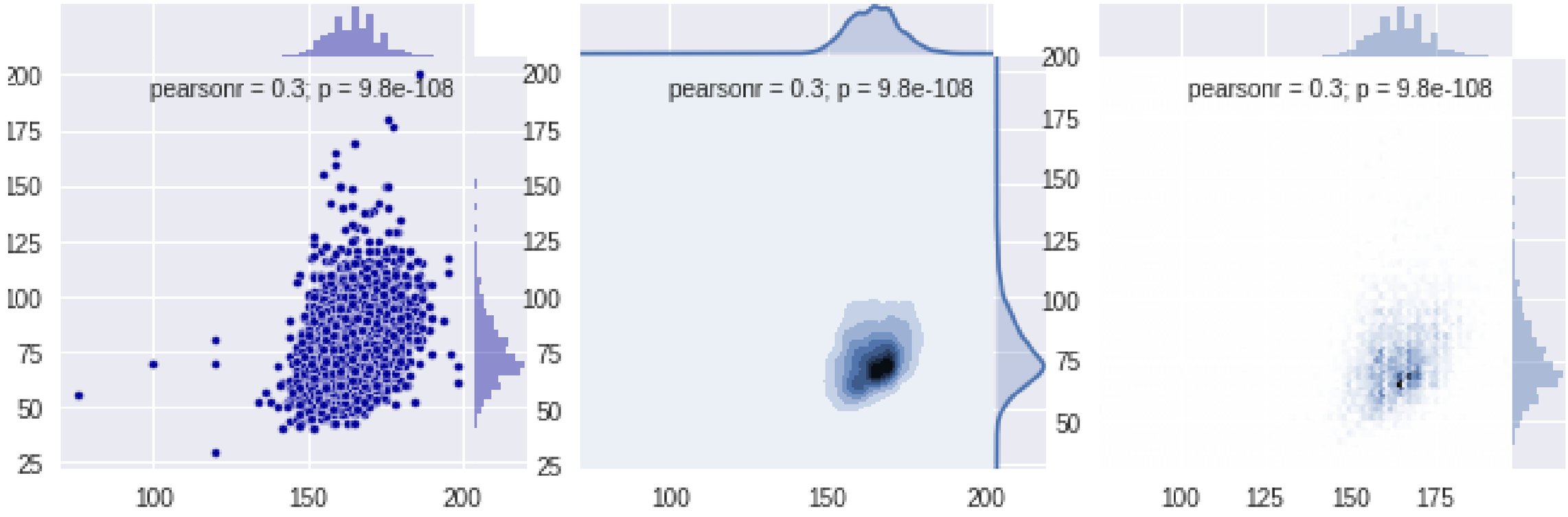
при правильном масштабе

кластеры

при правильном масштабе

Диаграмма рассеивания – лучший выбор

Задача о сердечно-сосудистых заболеваниях



признаки «рост-вес»

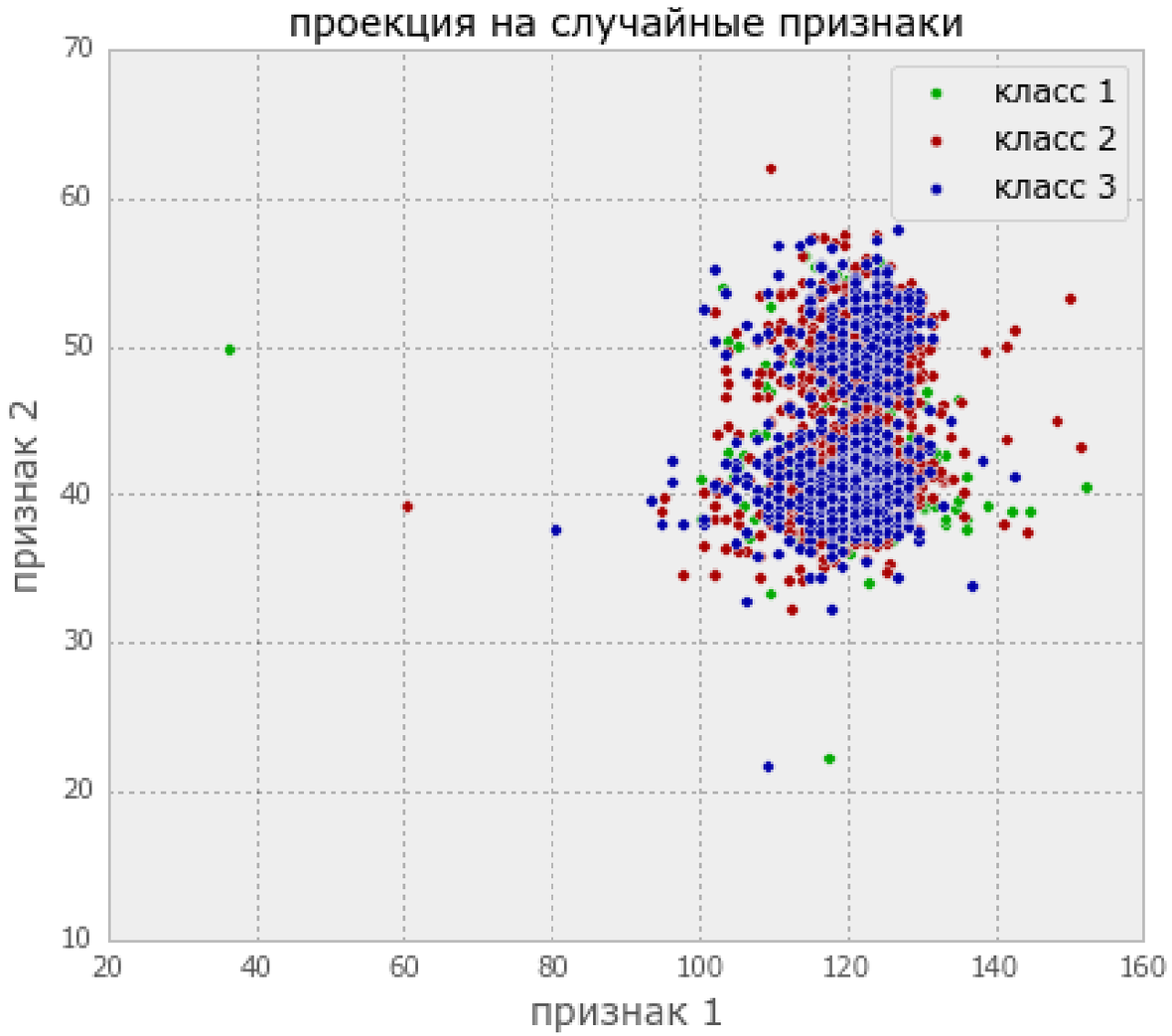
где видны выбросы?

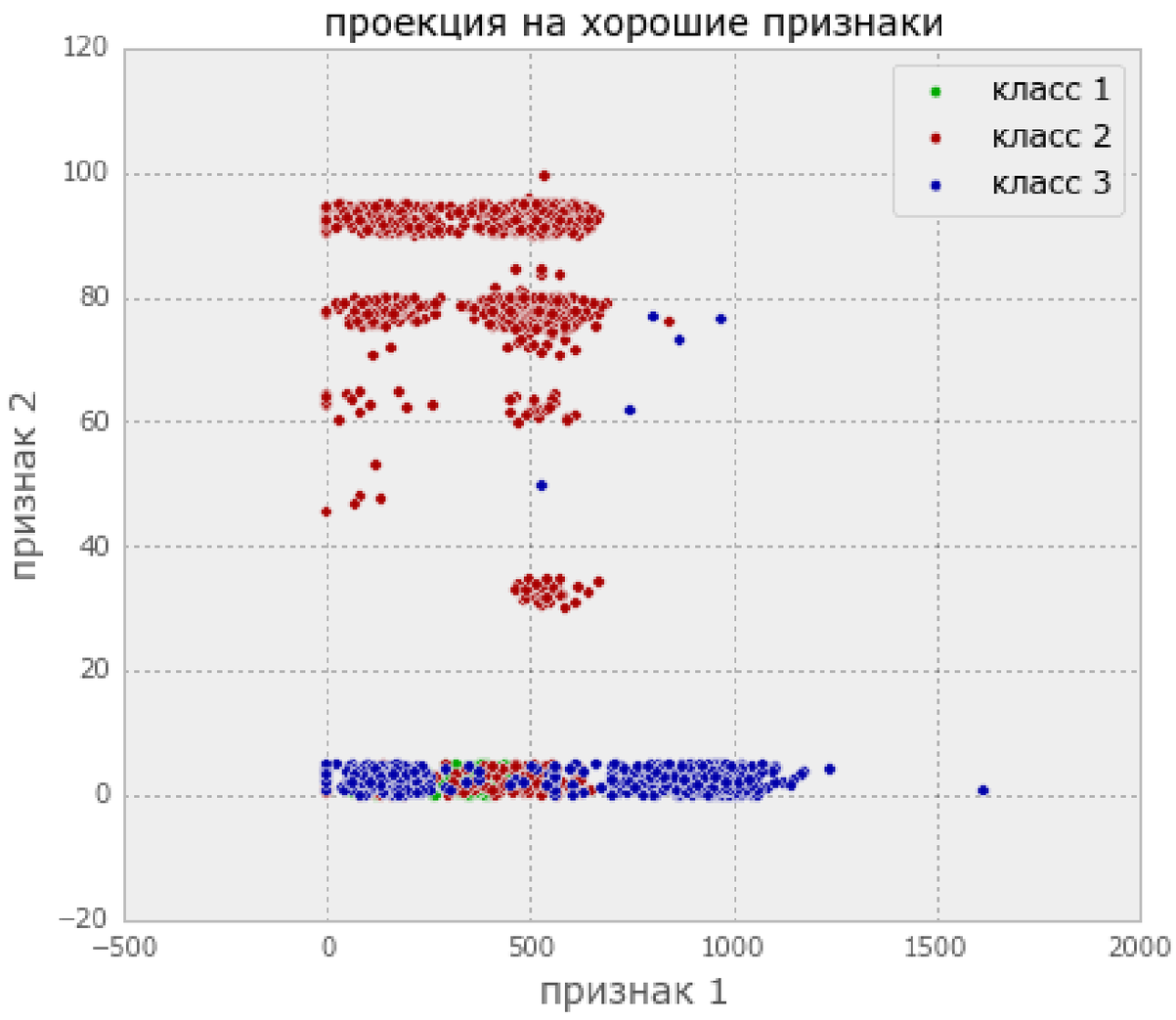
как сделать, чтобы и плотность анализировать?

Смотрим на пары признаков

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

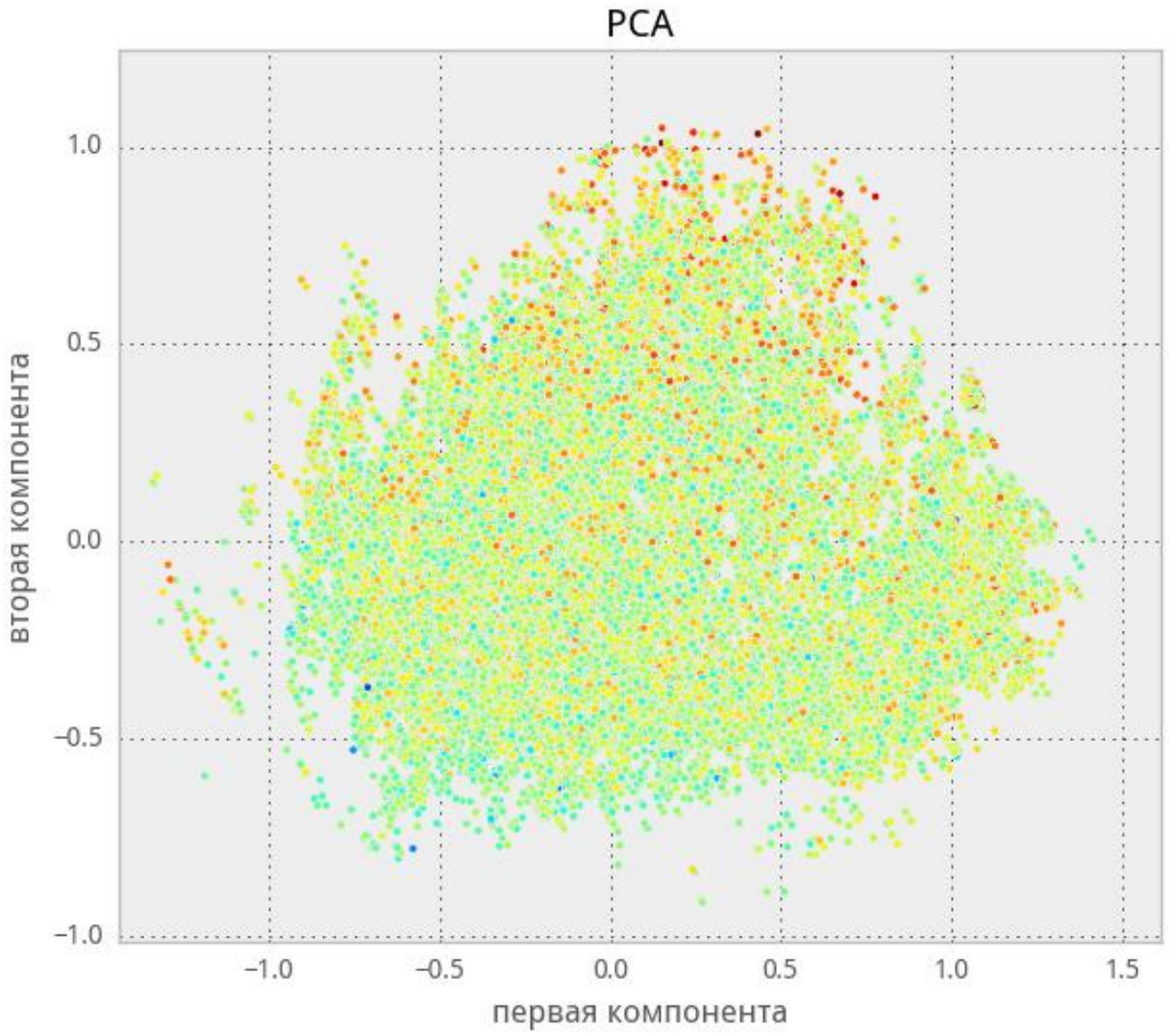
Смотрим на пары признаков

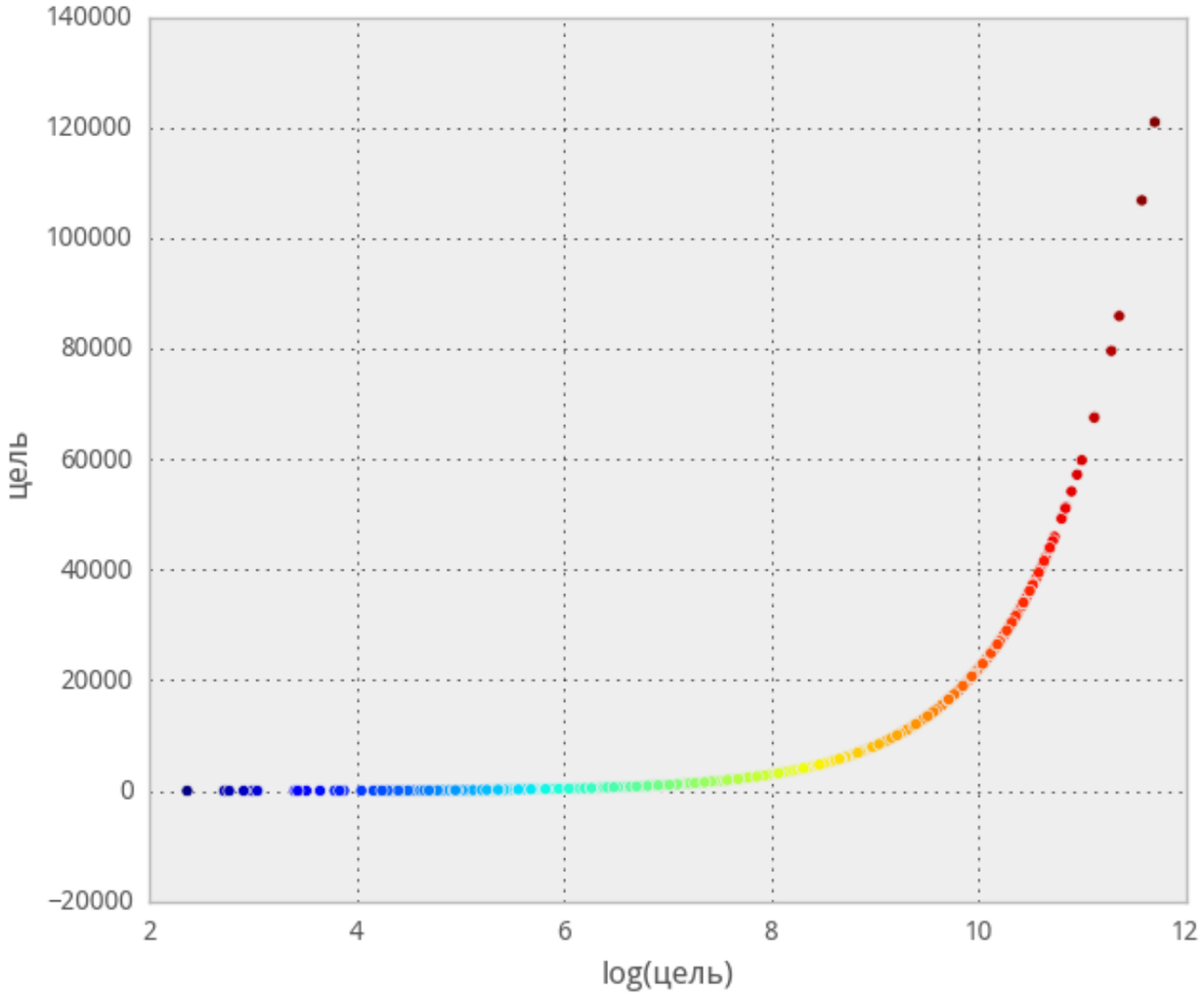




разница между случайными и хорошими признаками

Визуализация сгенерированных признаков «AllState»

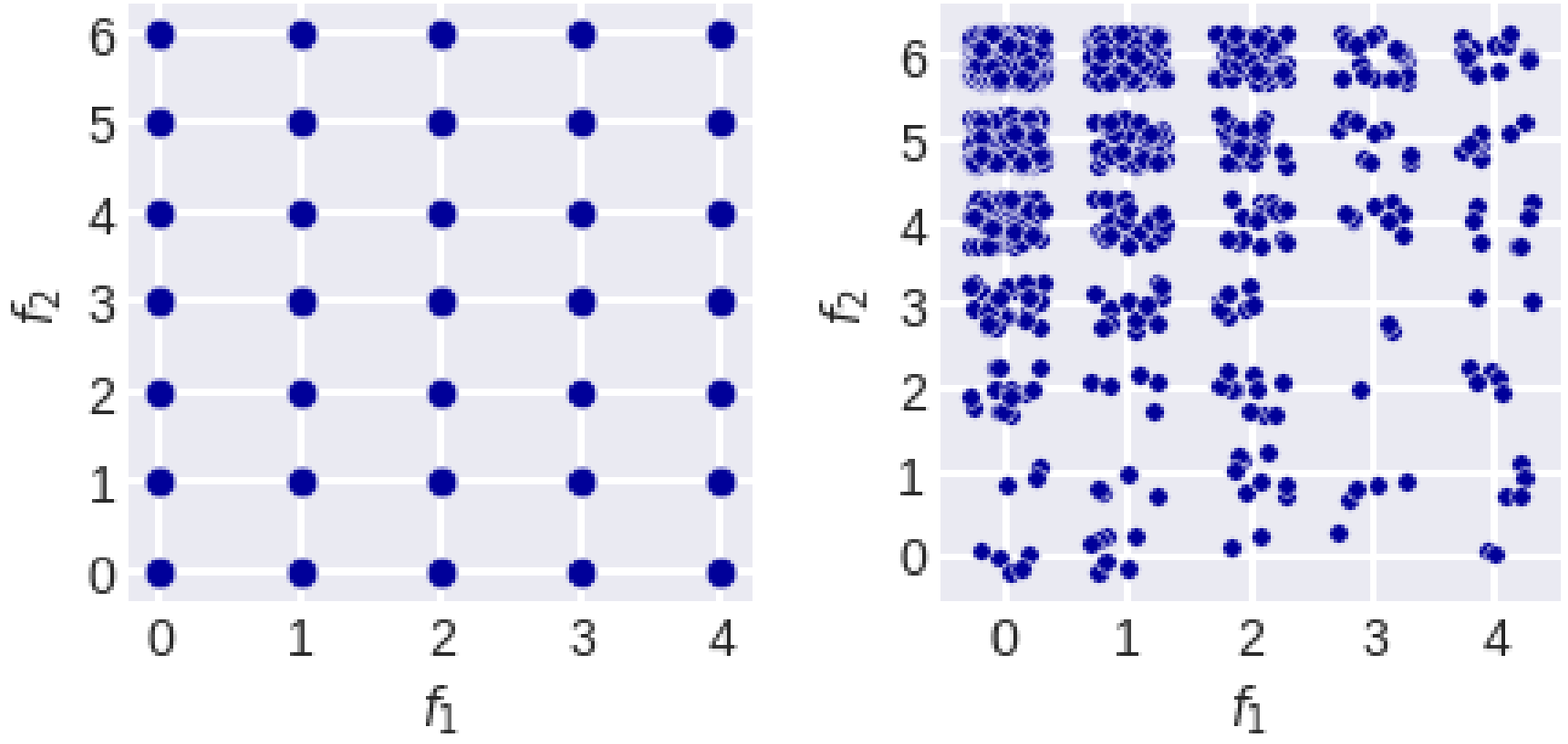




Что это за разложение / хорошее ли оно?

Диаграммы рассеивания дискретных признаков

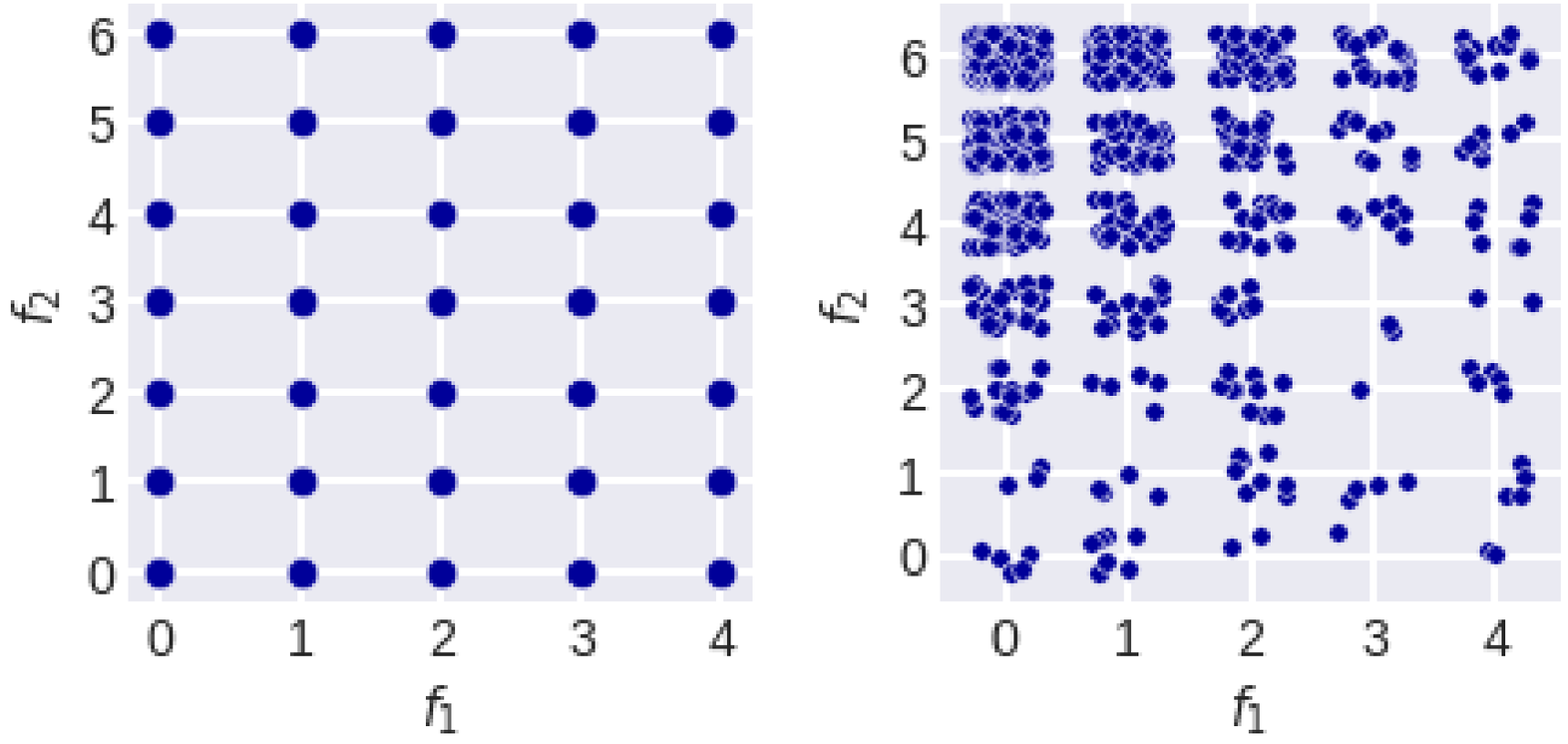
Зачем нужен Jitter



Что видно?

Диаграммы рассеивания дискретных признаков

Зачем нужен Jitter



Что видно?

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

Сводная таблица

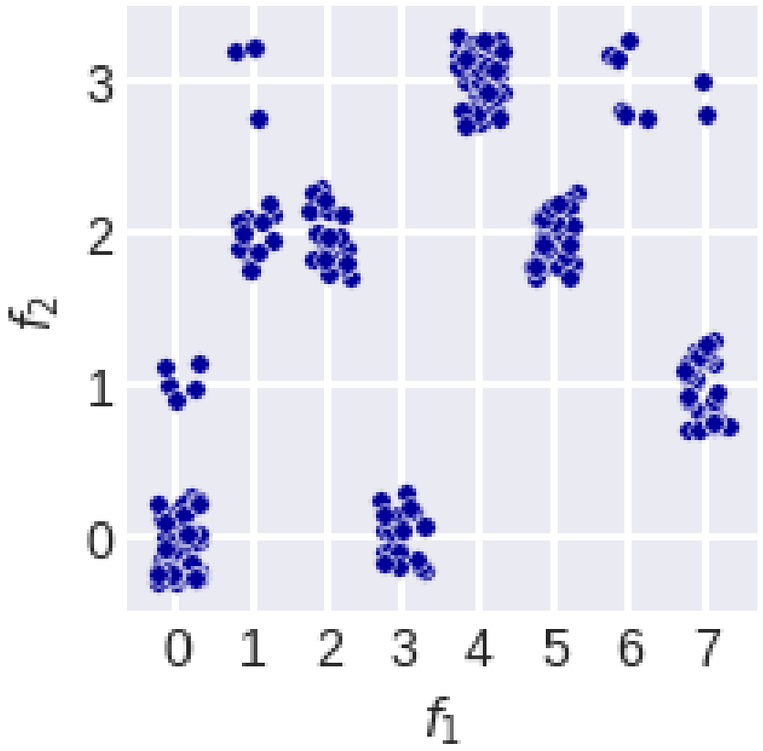
f_1 \ f_2	0	1	2	3	4	5	6
0	5	3	13	24	59	152	405
1	7	4	5	14	25	56	154
2	2	8	10	8	16	21	60
3	1	4	1	2	9	10	21
4	2	4	5	2	7	8	12

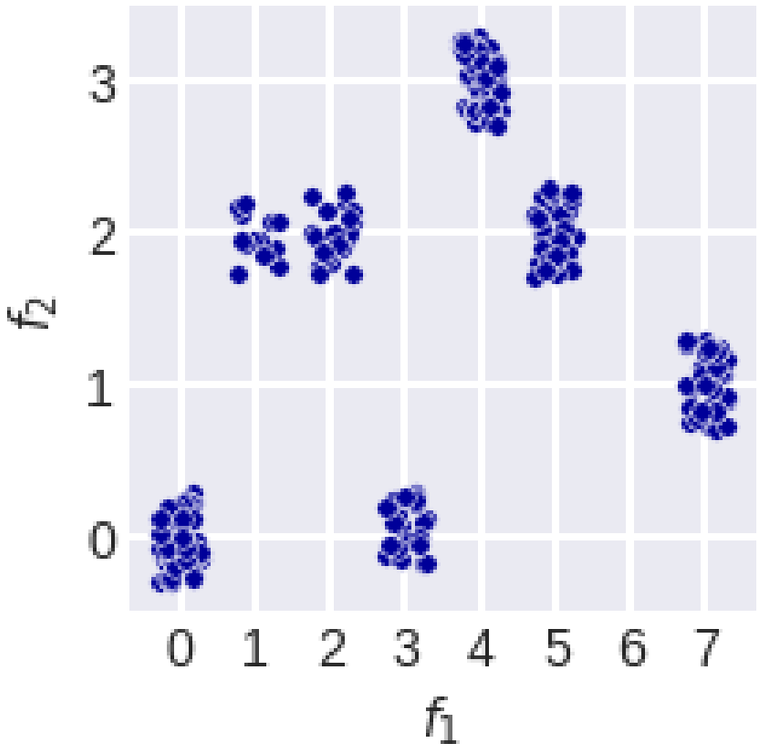
```
pd.crosstab(x1, x2)
```

Часто не нужно рисунков!

По таблице всё видно

Диаграммы рассеивания дискретных признаков

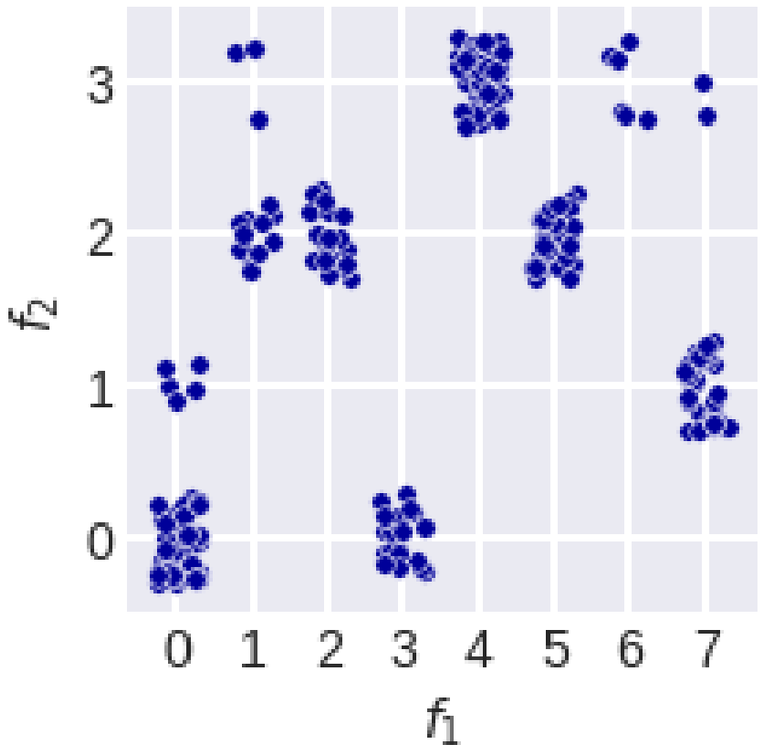


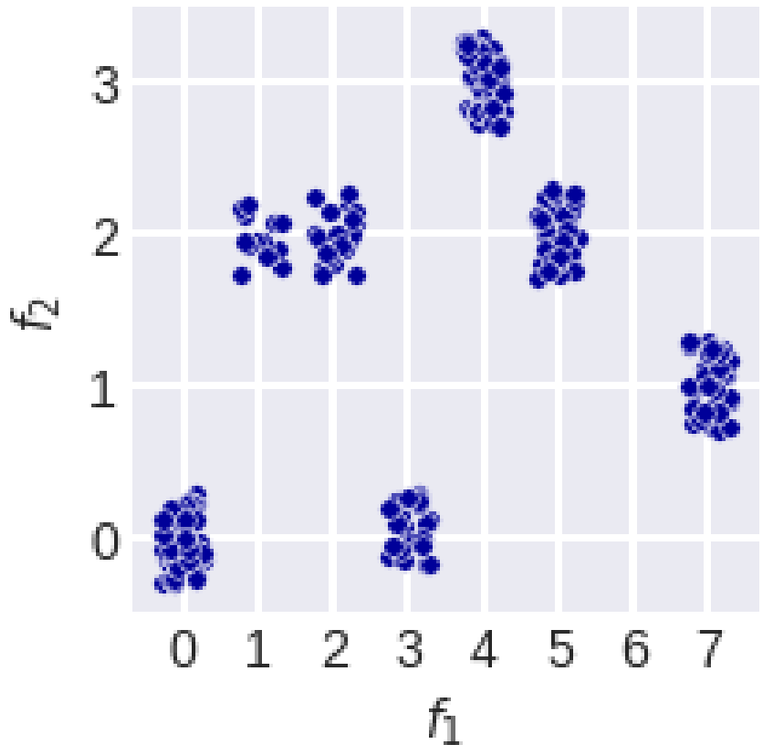


Справа – после удаления маленьких кластеров!

Что здесь видно?

Диаграммы рассеивания дискретных признаков





Один признак – уточнение другого!

Как это использовать?

«Liberty»

Из задачи «Liberty»

Верхняя треугольная Зависимость

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

V6 / V14

Обоснование необходимости использования пар признаков

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

V11 / V13

	A	B	C	D	E
N	3.88	5.10	4.57	5.52	3.95
Y	3.81	4.32	4.23	4.18	3.94

mean target V11 / V13

```
df.groupby(['x1', 'x2'])['target'].mean().unstack('x2')
```

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,

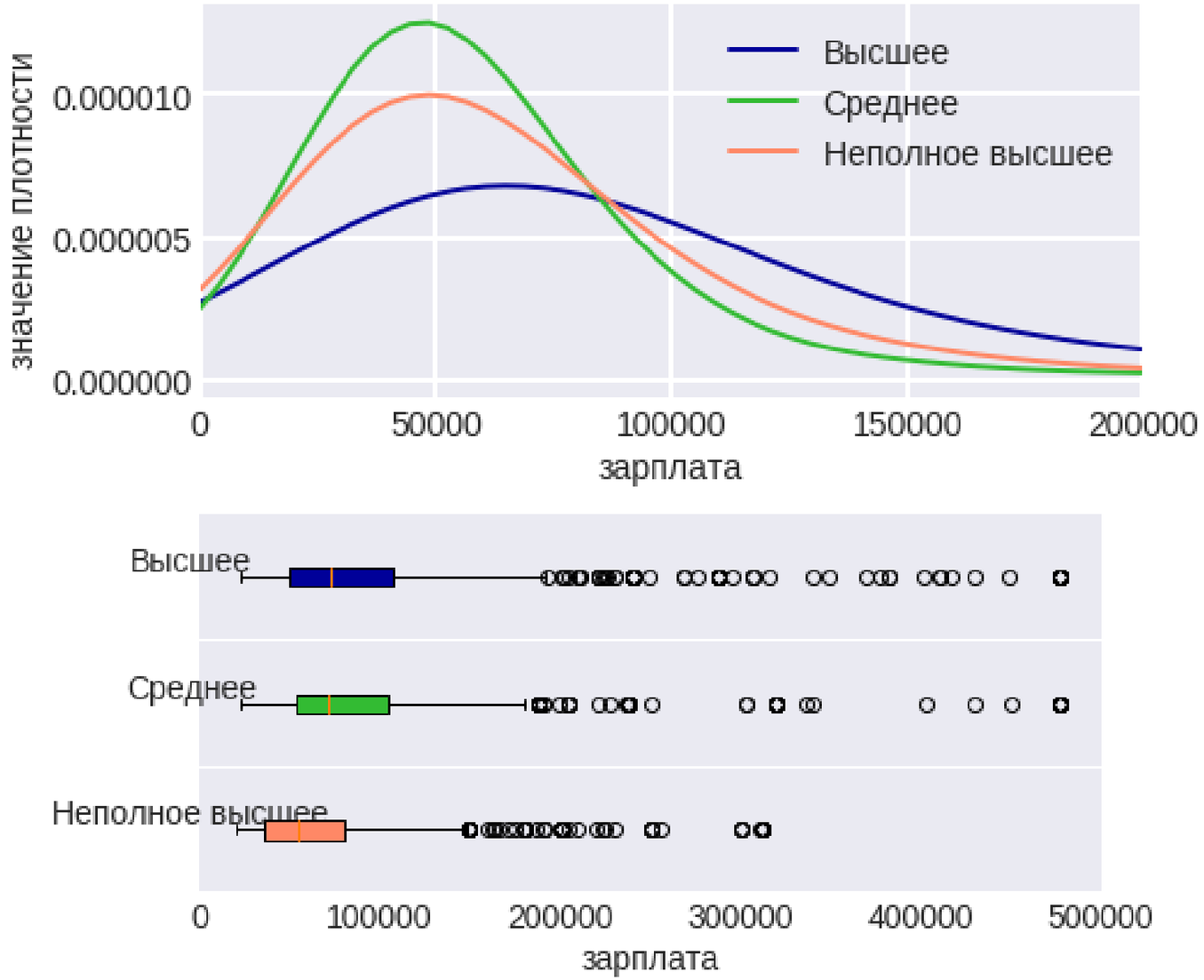
что один категориальный признак – уточнение другого

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

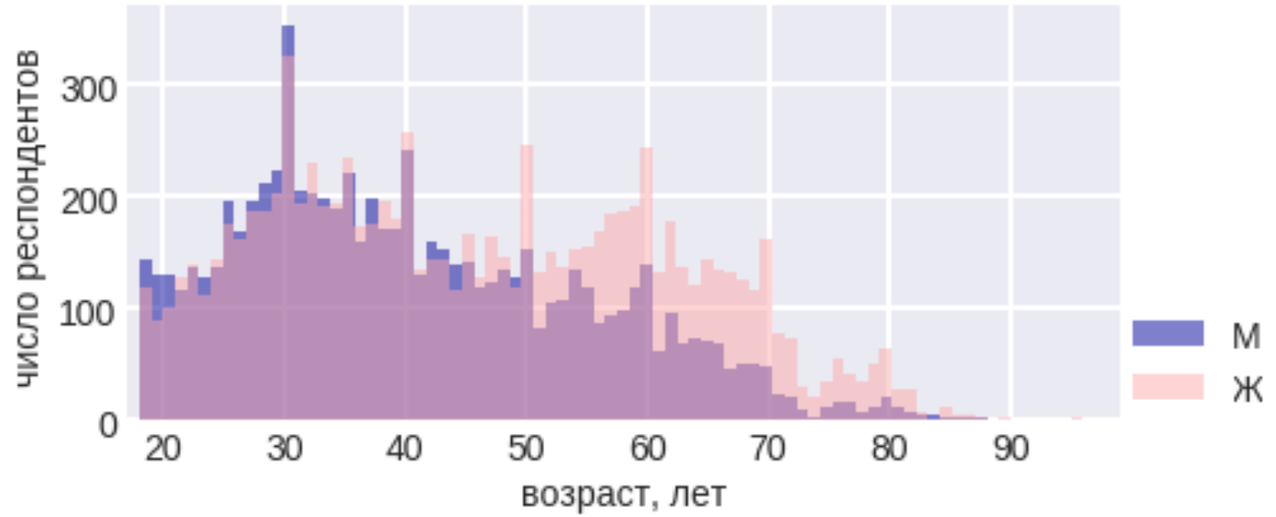
char_2	type 1	type 2	type 3
char_1			
type 1	15251	0	0
type 2	0	77314	96553

Как использовать это знание?

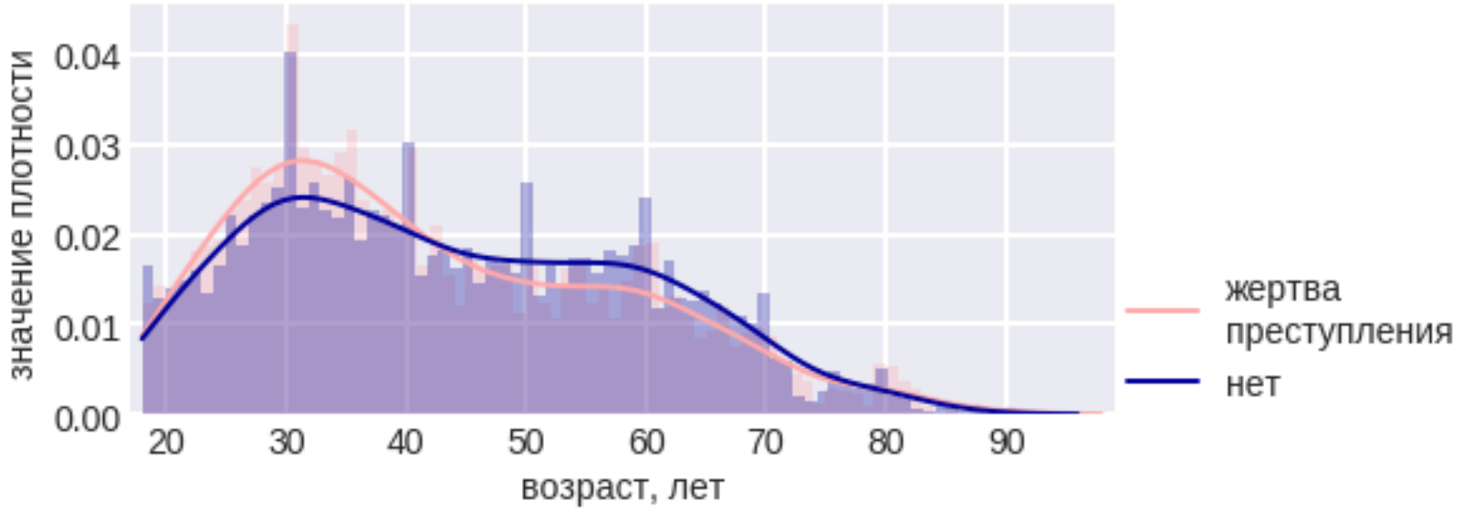
Пара «вещественный признак – категориальный»



Пара «вещественный признак – категориальный»

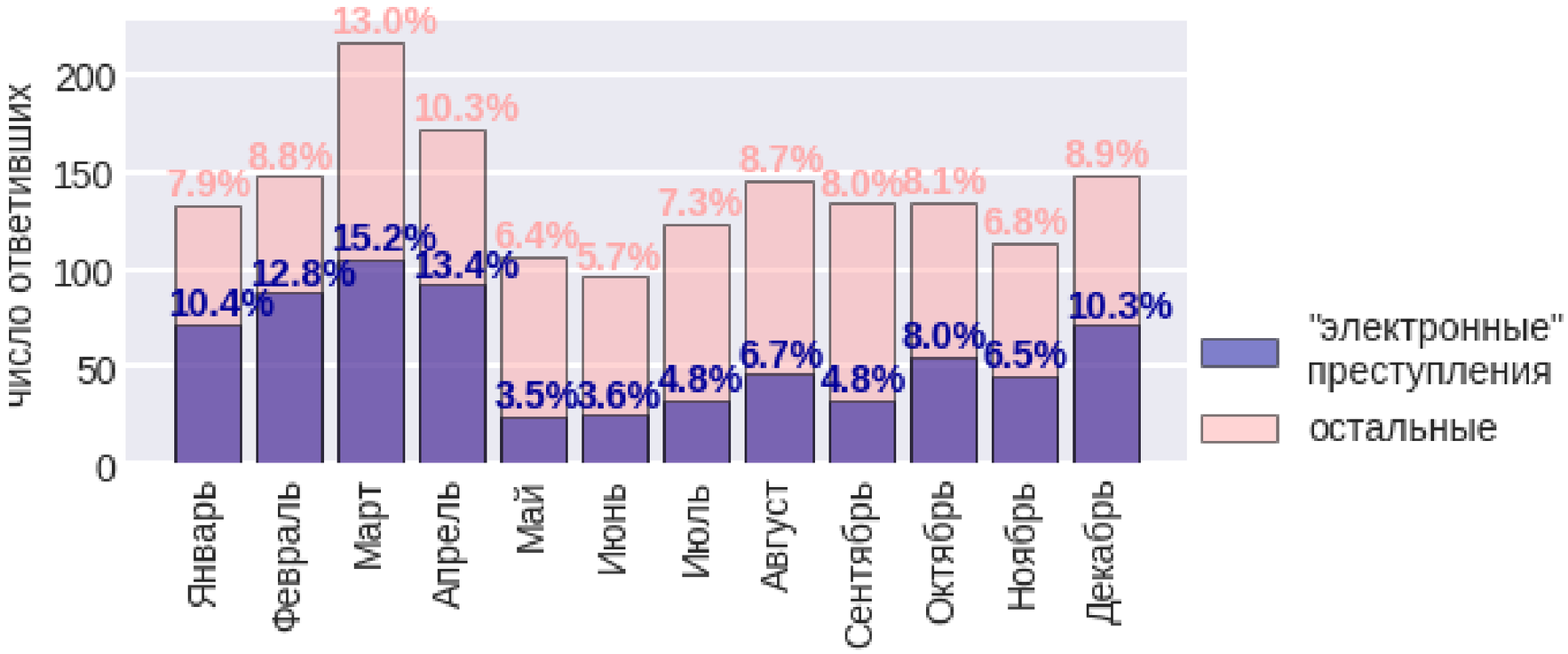


Распределение респондентов по возрасту и полу



Распределение возрастов жертв преступлений и остальных респондентов

Пара «бинарный признак – категориальный»

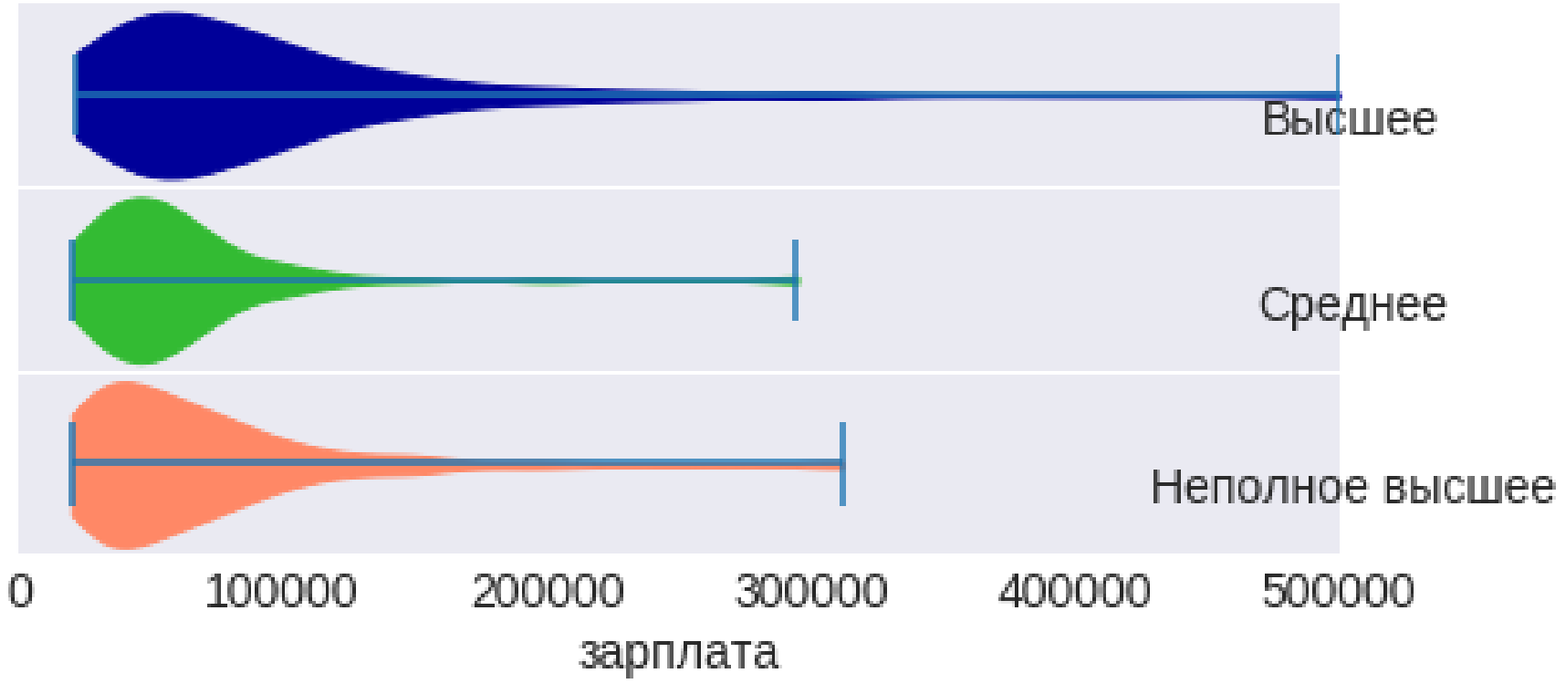


Здесь наоборот – по категориям средние значения бинарного

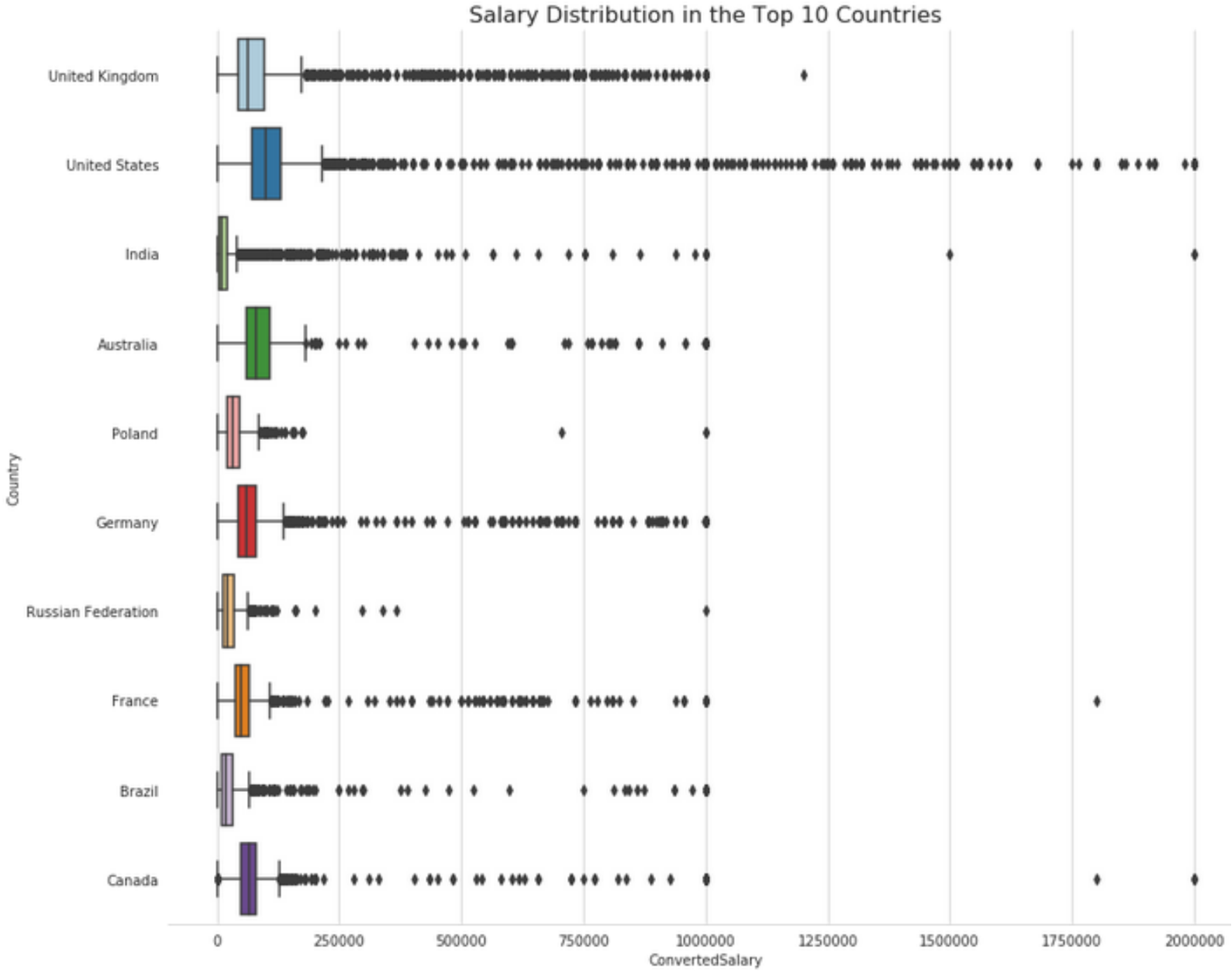
показан даже 3й признак – вид преступления

что видно из рисунка? какие выводы можно сделать?

Всё это не очень наглядно...

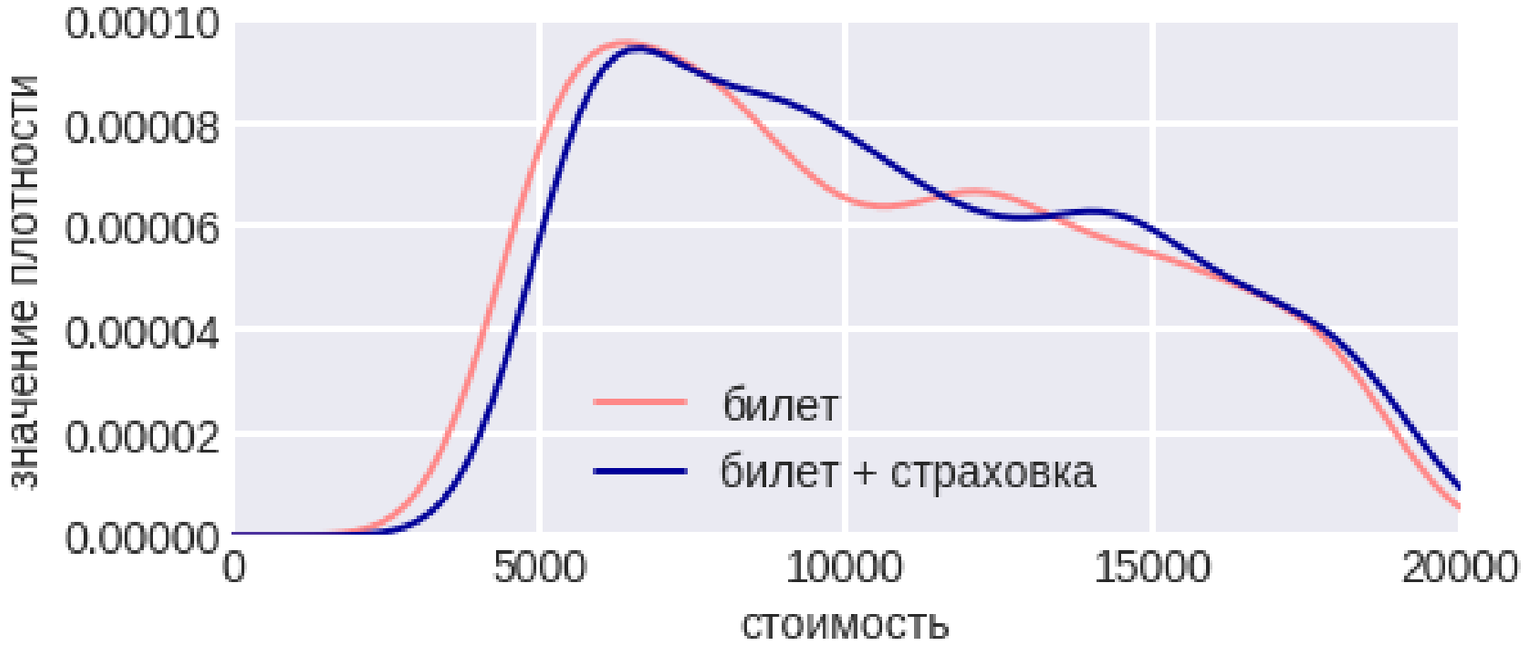


Пример использования «ящика с усами»



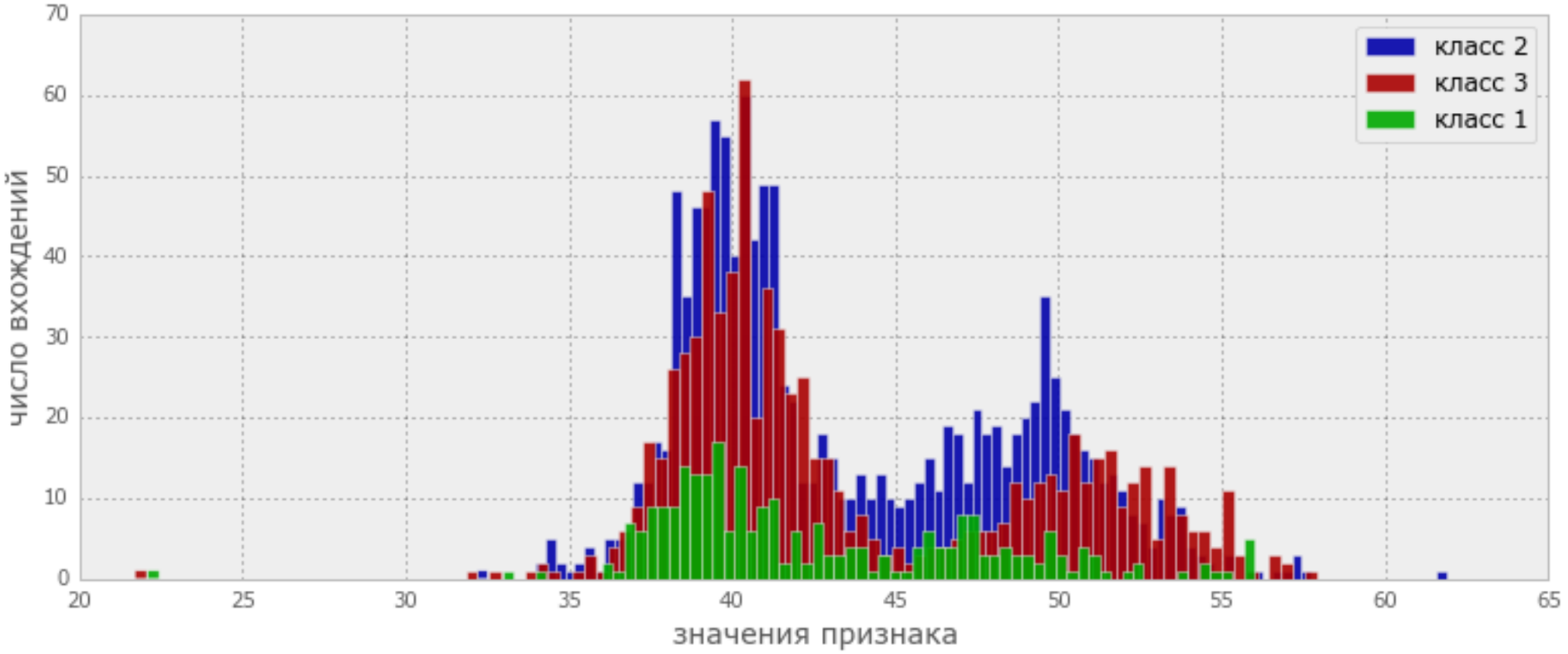
https://www.kaggle.com/djaballah/stackoverflow-beginner-eda

Задача «Ozon Travel»



Всегда ставьте под сомнение свои выводы!

Как распределена цель на признаках

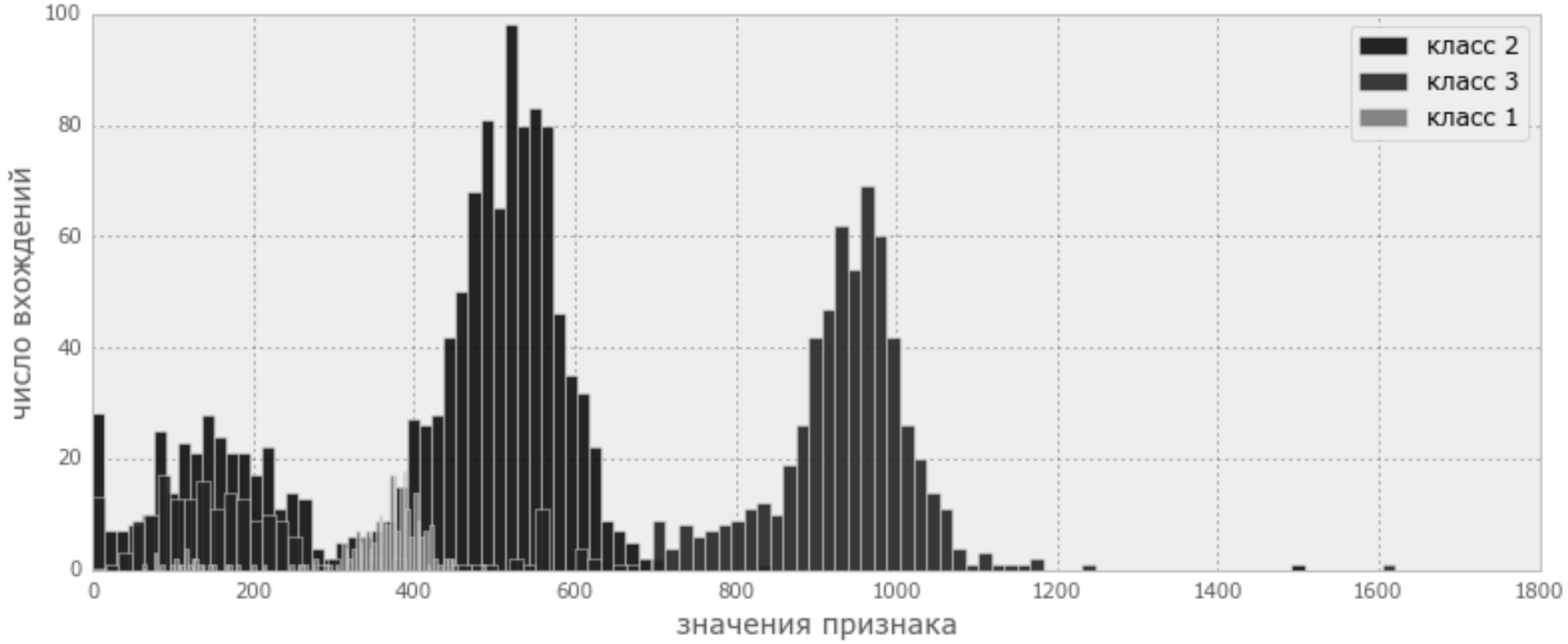


Как распределена цель на признаках

Чем плох рисунок?

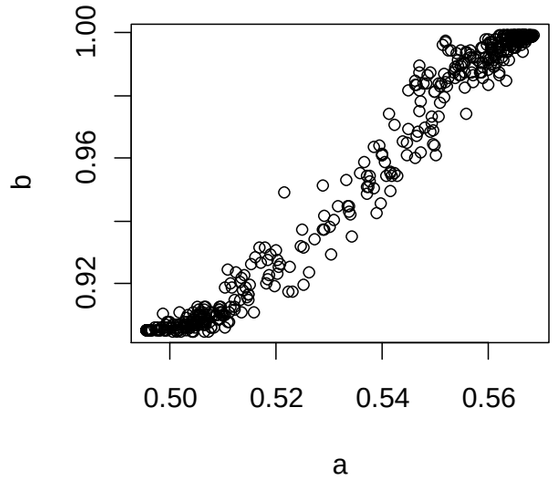
Чем признак отличается от предыдущего?

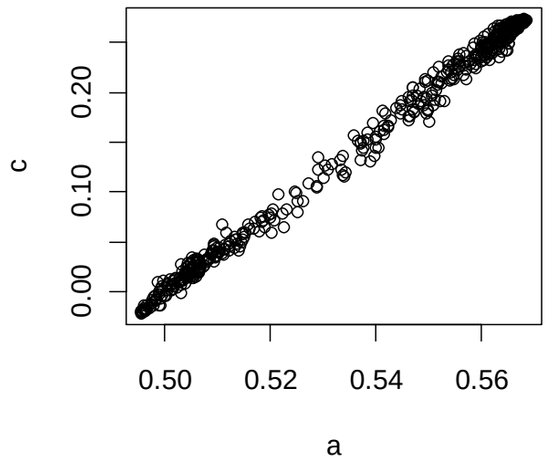
Как распределена цель на признаках

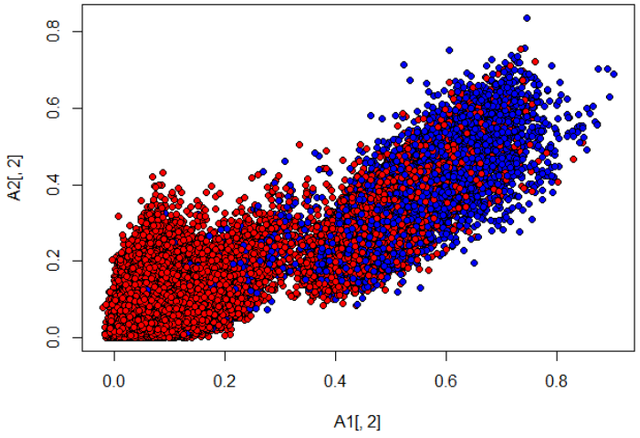


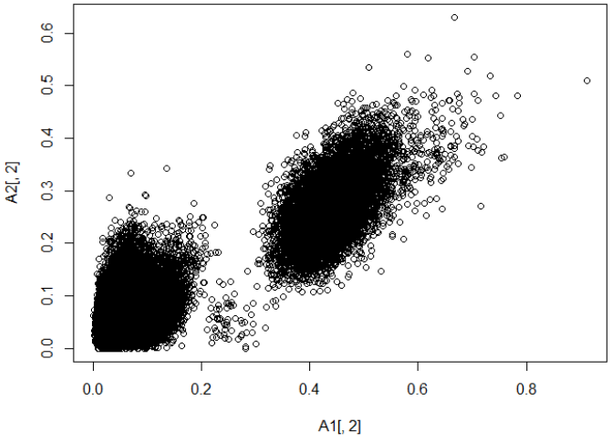
Вот чем...

Визуализация ответов двух алгоритмов: как найти ошибку используя бенчмарк



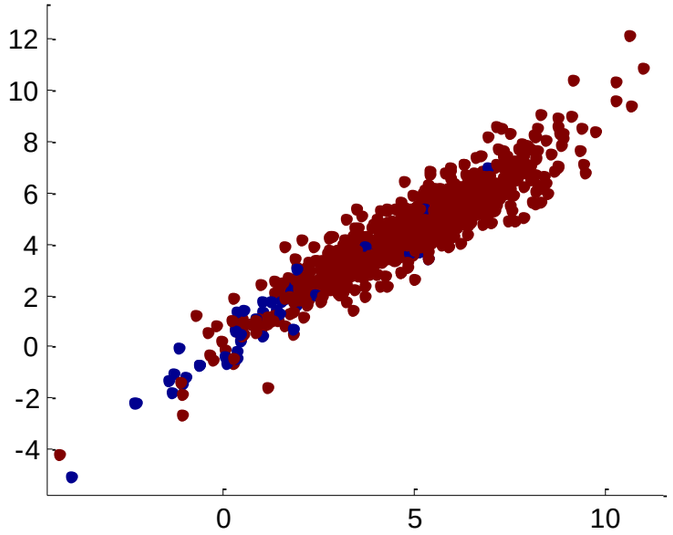




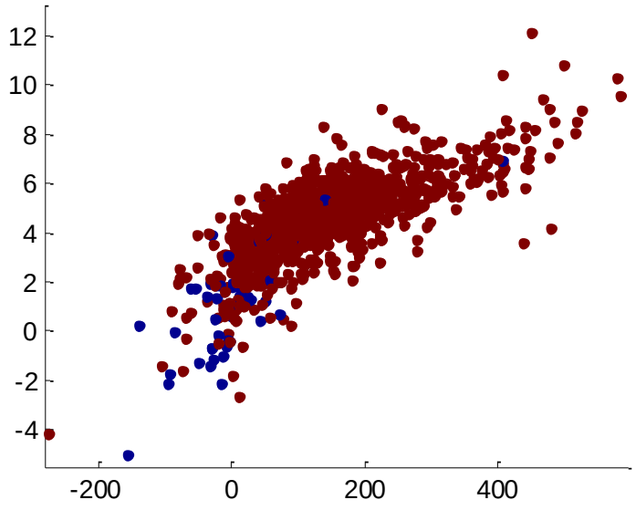


Совет: создавайте бенчмарк!

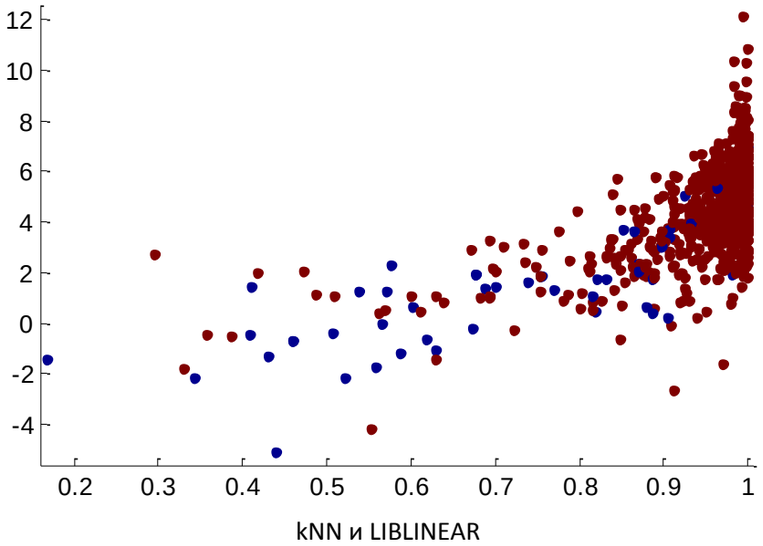
Ещё о визуализации «алгоритм-алгоритм»



Разные LIBLINEAR



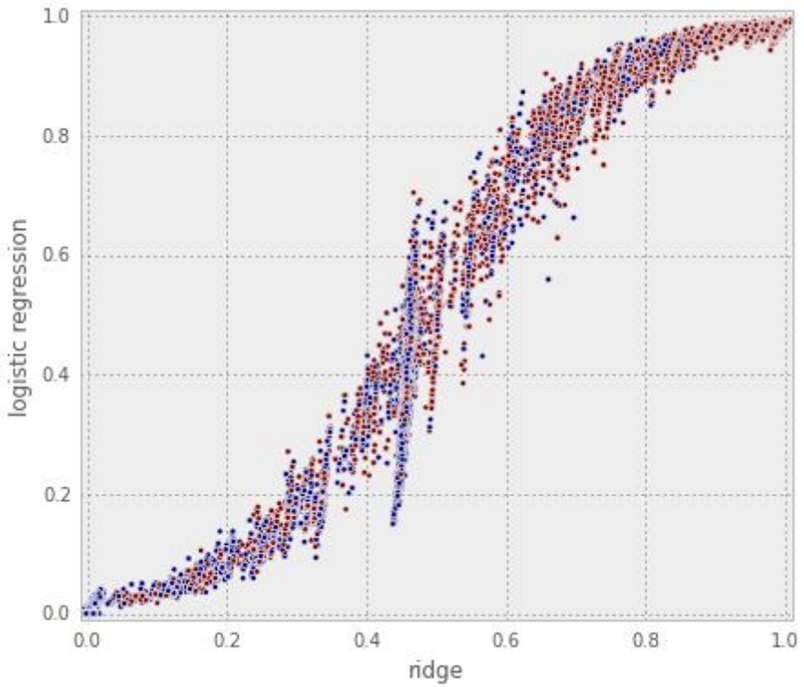
LIBLINEAR и PERCEPTRON

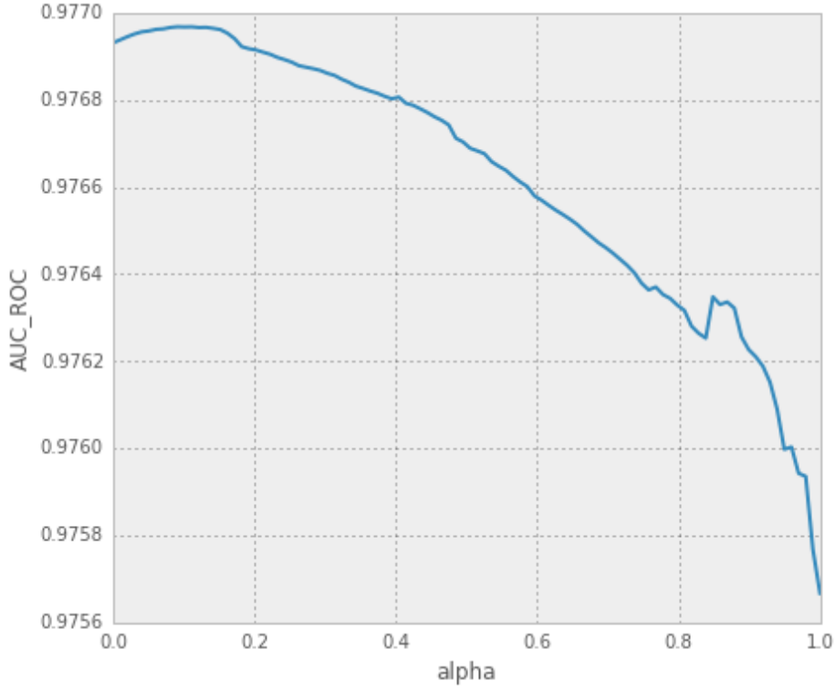


kNN и LIBLINEAR

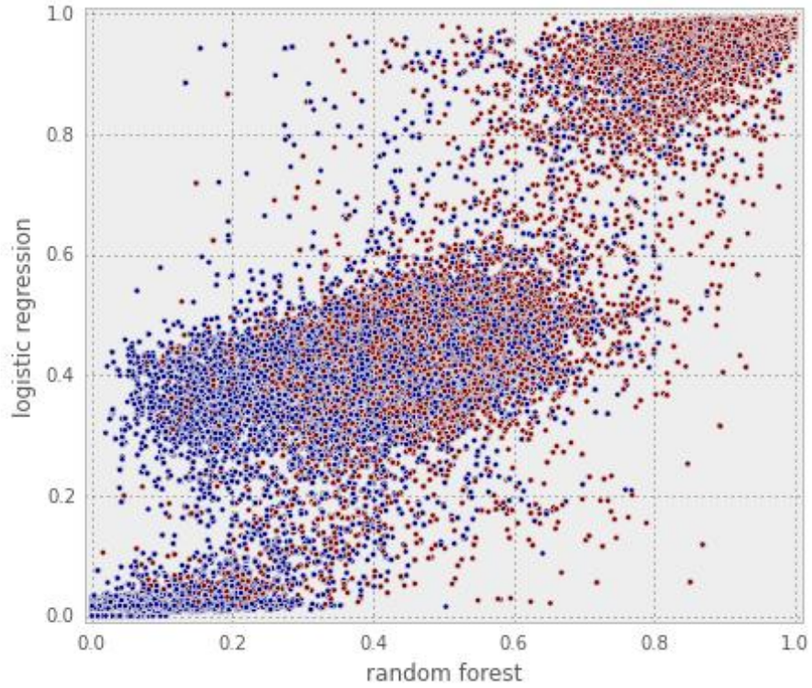
В задаче AMAZON

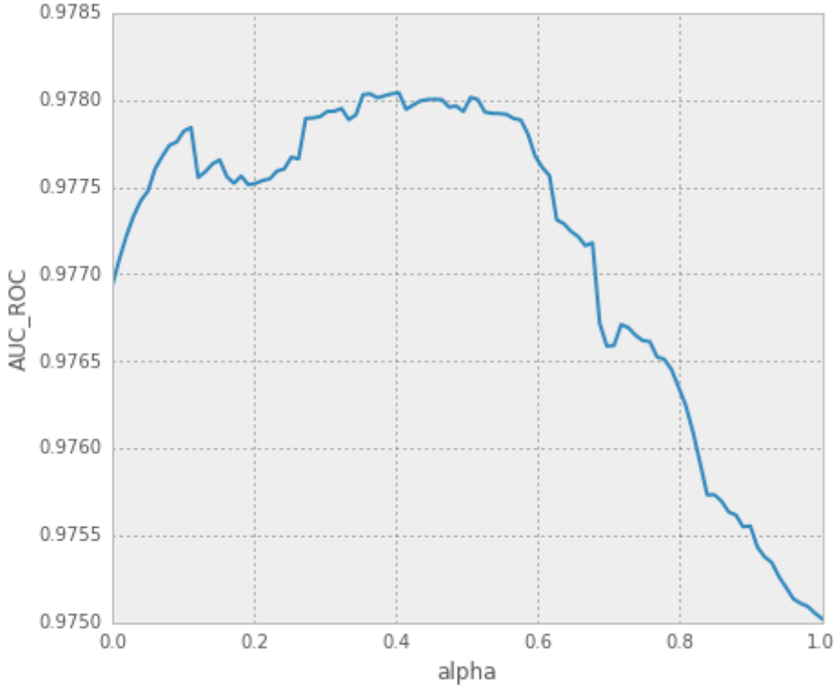
Ансамбль регрессия + логистическая регрессия



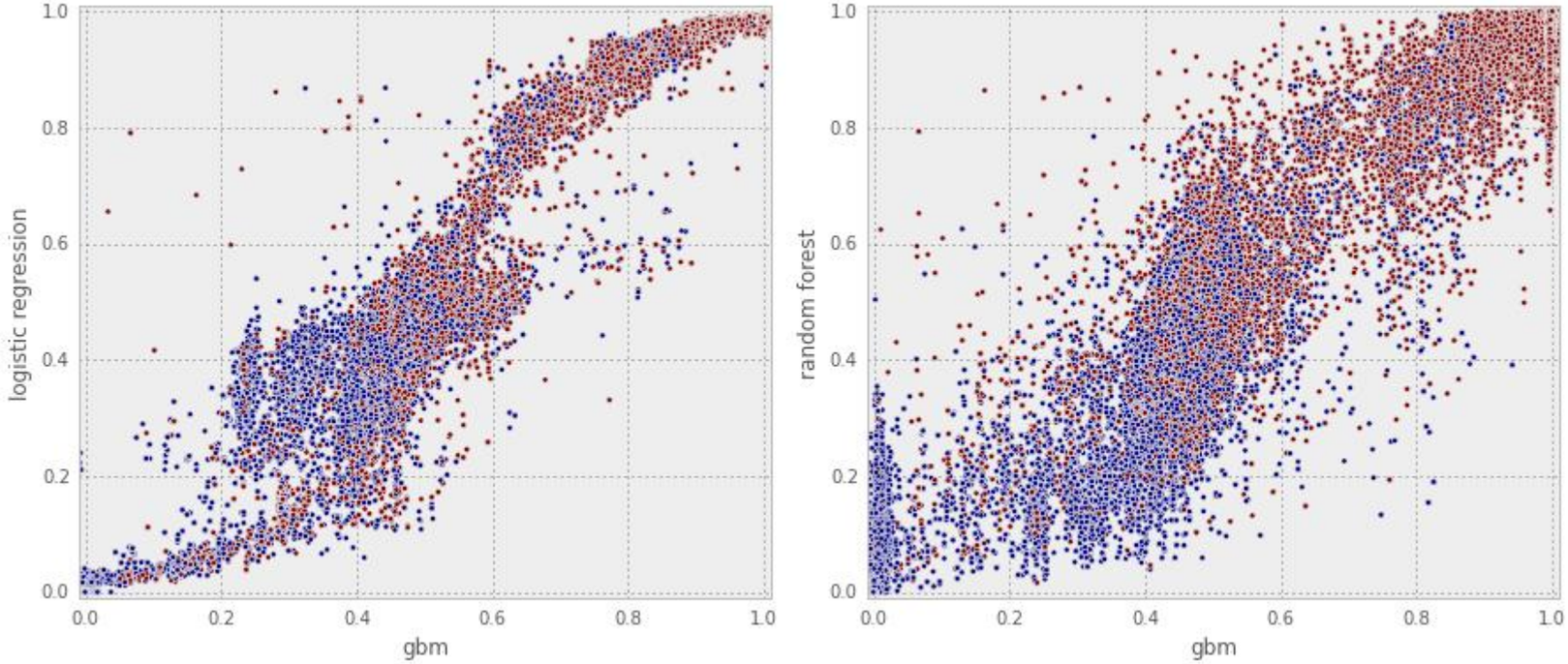


Ансамбль случайный лес + логистическая регрессия

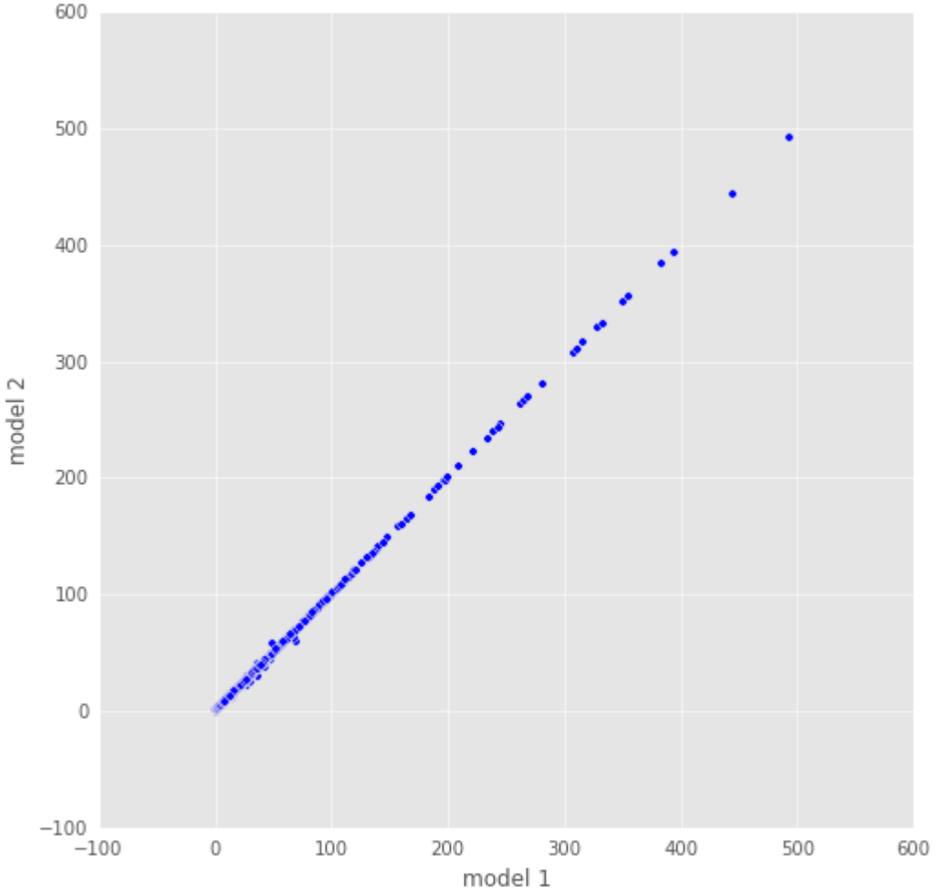




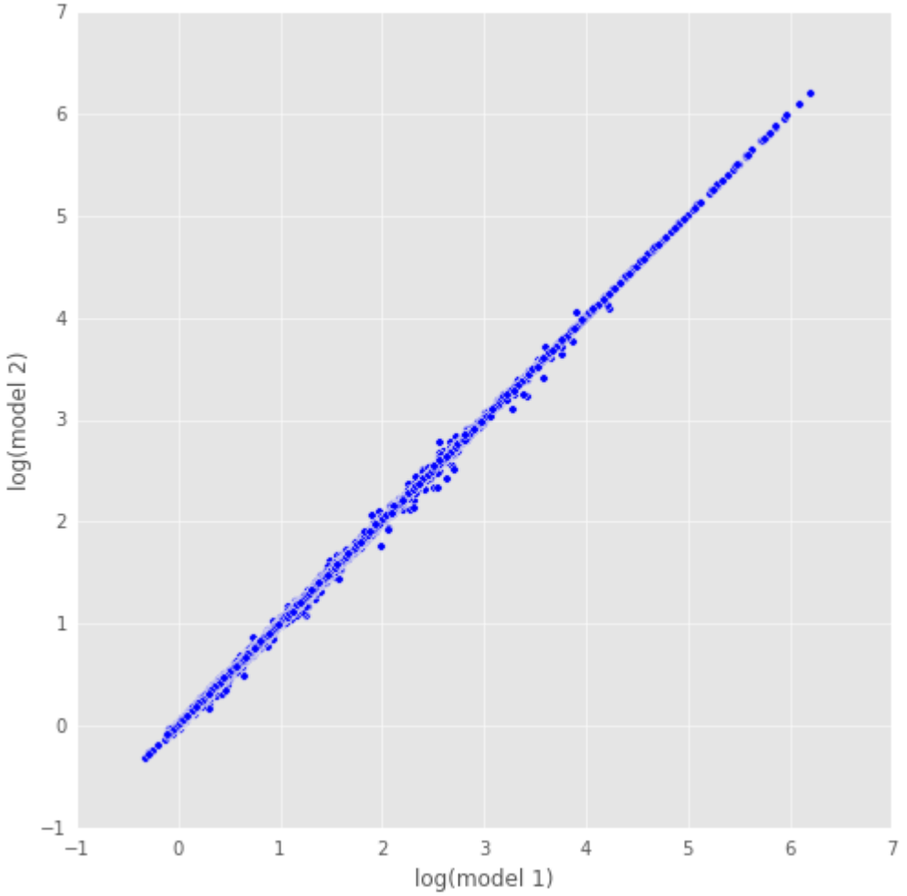
Ансамбли с gbm



Ещё о визуализации «алгоритм-алгоритм»

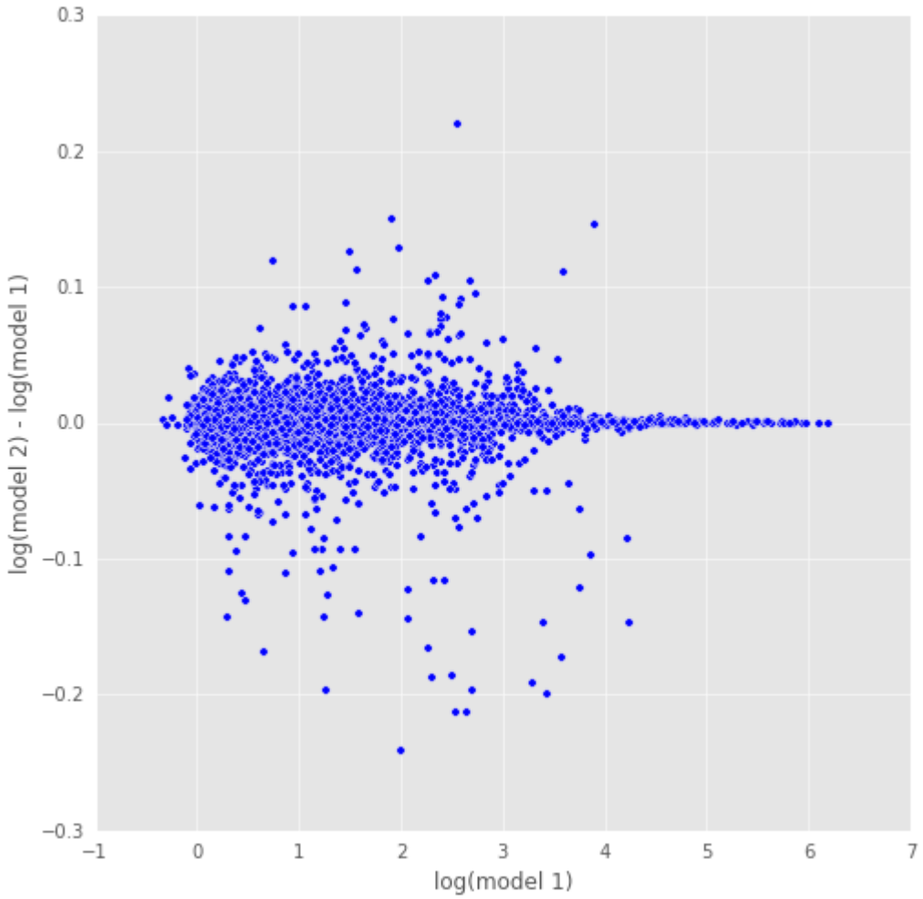


Две модели



Опять логарифмирование шкал!

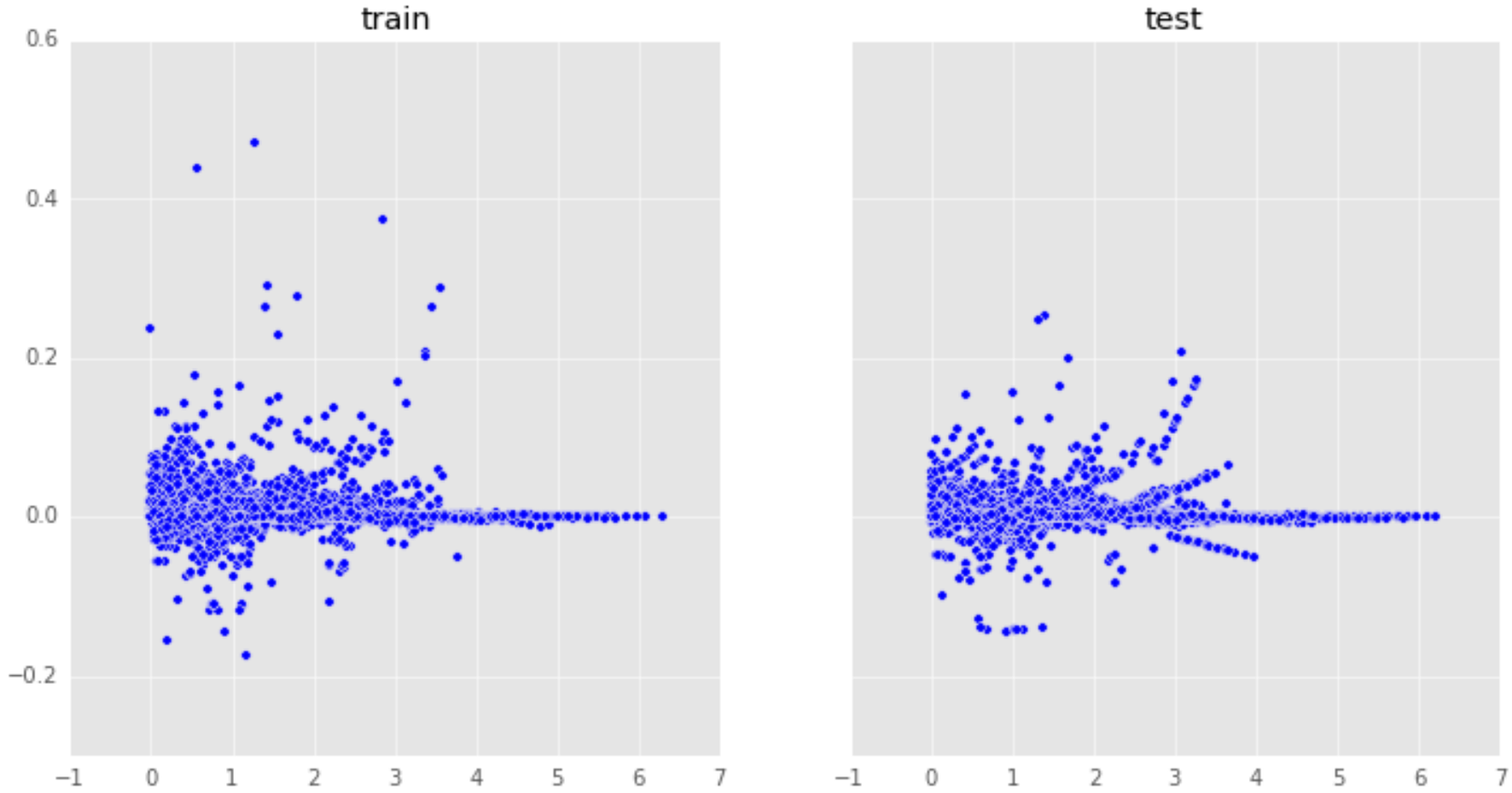
Ещё о визуализации «алгоритм-алгоритм»



Опять смотрим разницу ответов

Наблюдение: при больших значениях модели работают идентично!

Ещё о визуализации «алгоритм-алгоритм»

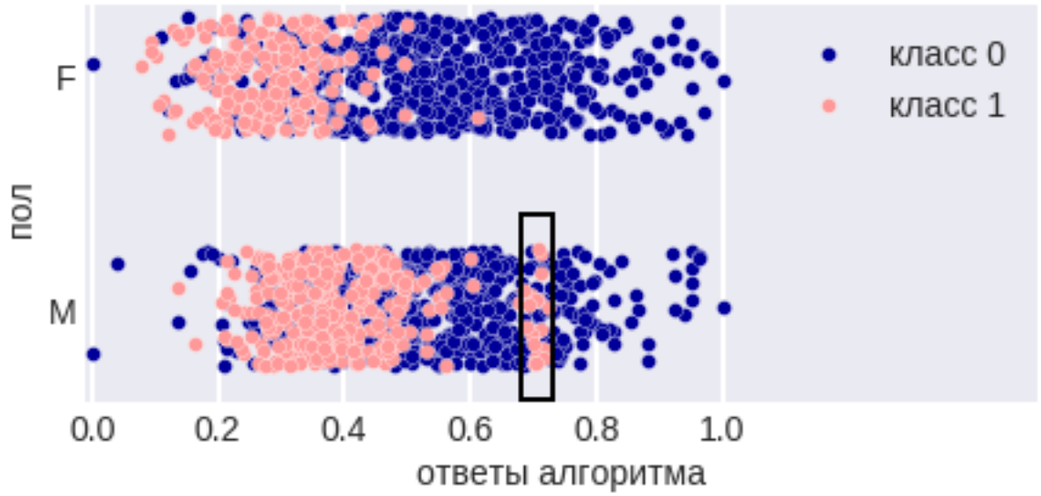


На контроле подозрительные линии…

Что это может значить?

Что делать?

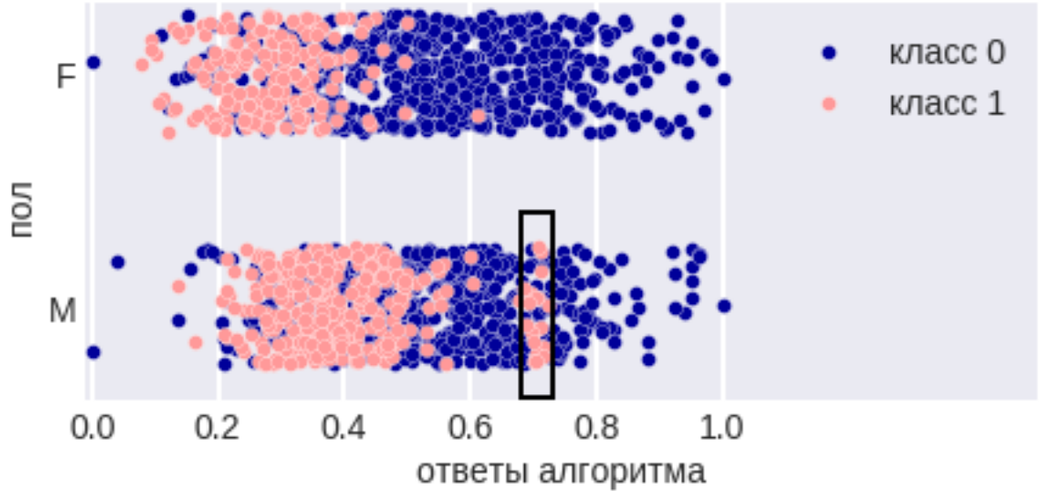
Ответы алгоритма – признак



Что видно?

Задача «~Analytics»

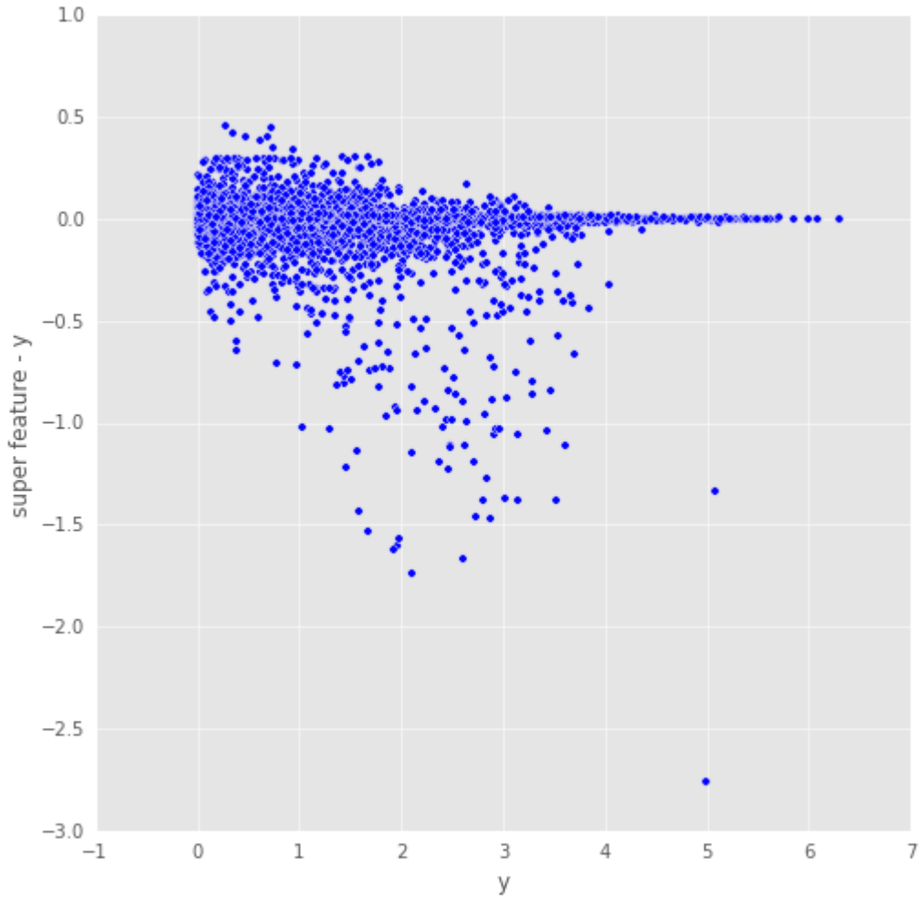
Ответы алгоритма – признак



Что видно:

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

Residual plot

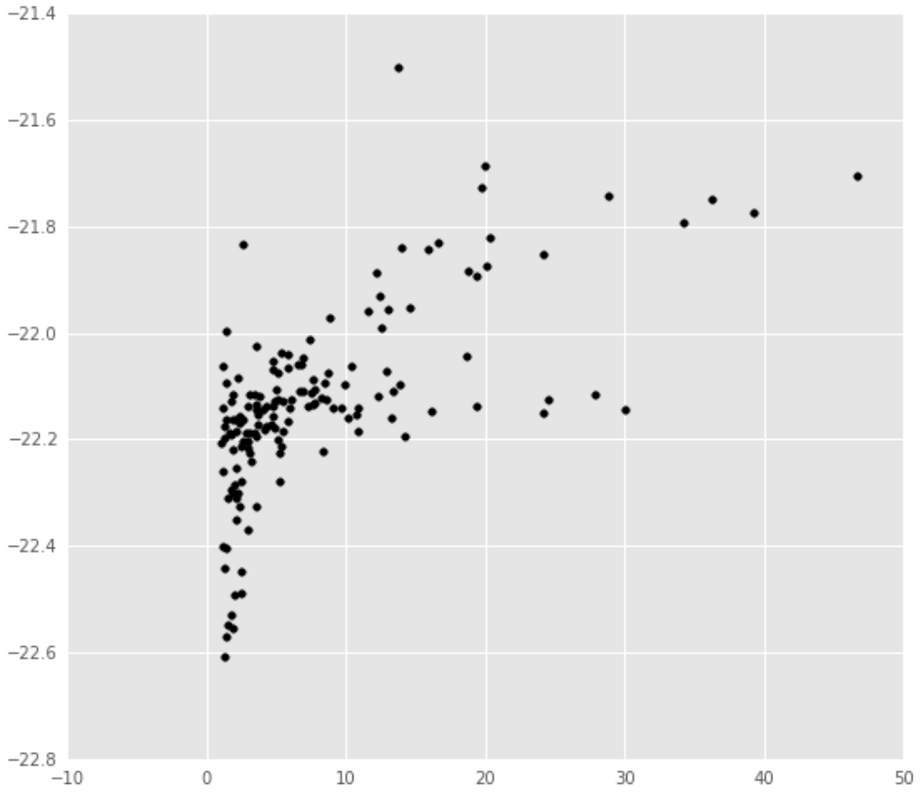



```
plt.scatter(np.log(y2), np.log(train2.mnk.values) + train2.tmp.values - np.log(y2))
```

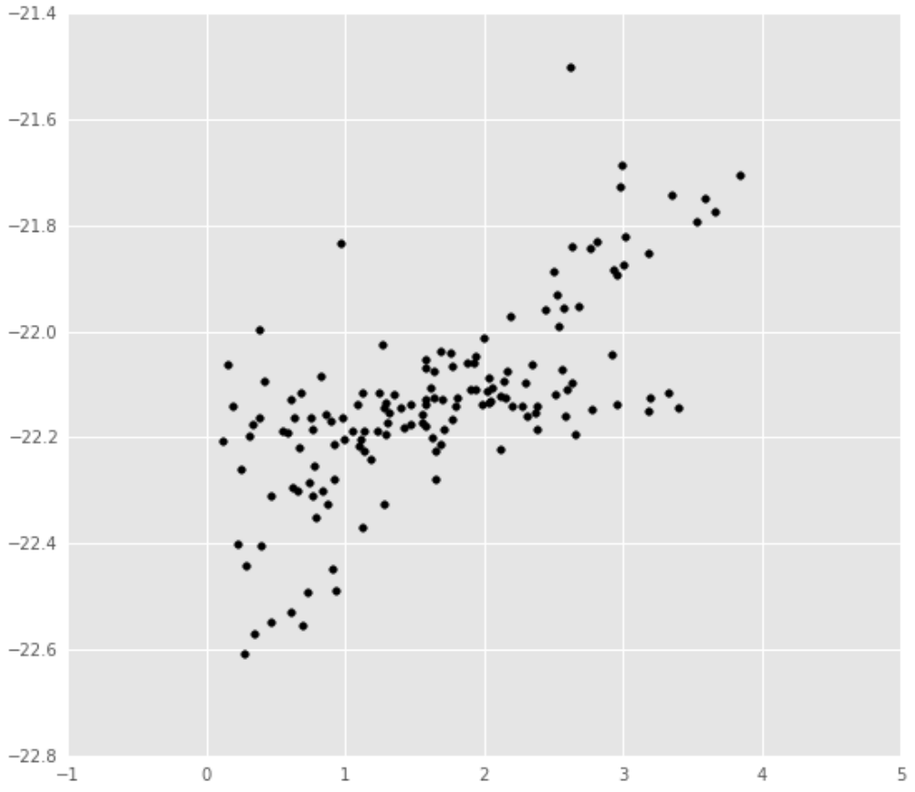
диаграмма рассеивания «невязка» – «прогнозируемая величина»

могут подсказать нужную трансформацию

Необходимость логарифмирования можно не заметить на меленьких выборках



До логарифмирования



после

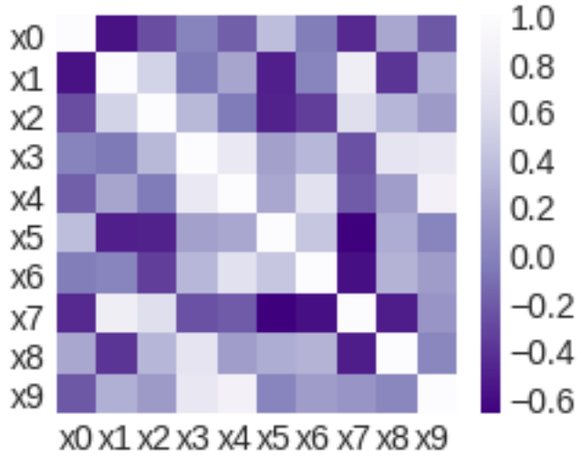
Корреляция между признаками

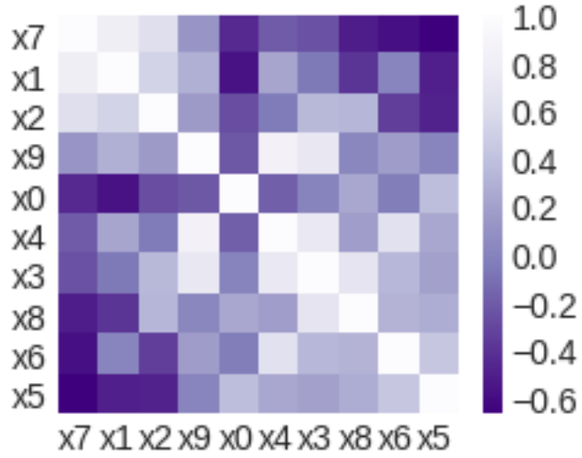
$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{m}(x_i - \mathrm{mean}(X))(y_i - \mathrm{mean}(X))}{m-1}$$

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

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

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

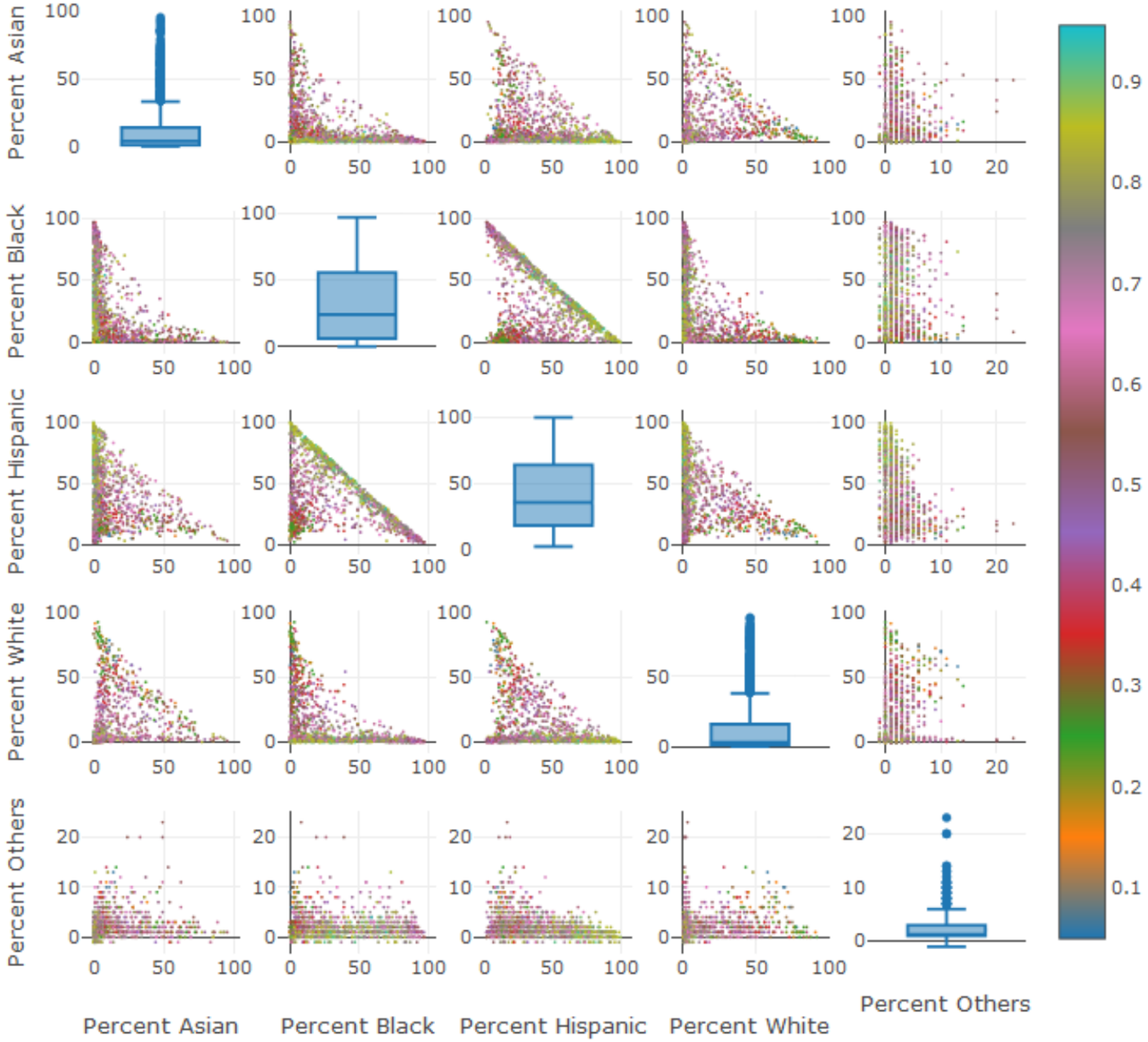
Корреляция между признаками

Корреляция – линейная зависимость…

Можно

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

Информация по всем парам – как правило, сильно перегружена



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

3D-визуализации

Третий признак

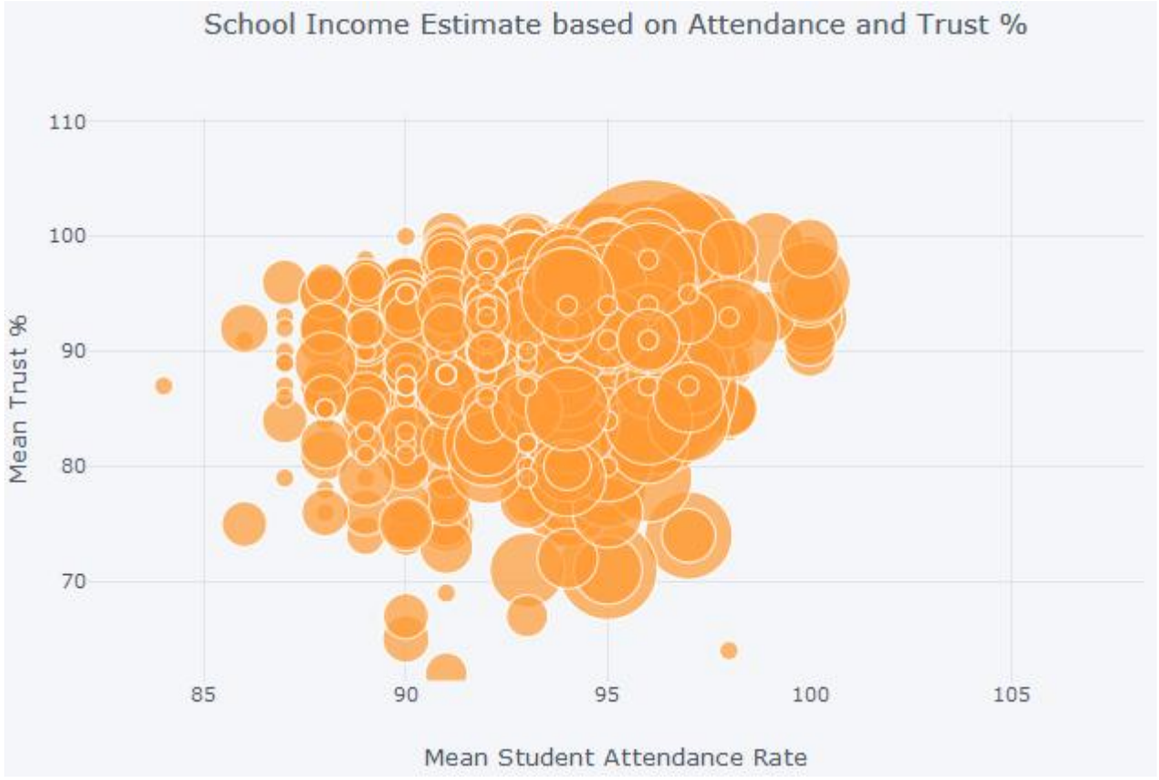
- **цвет**
- **размер**
- **форма**

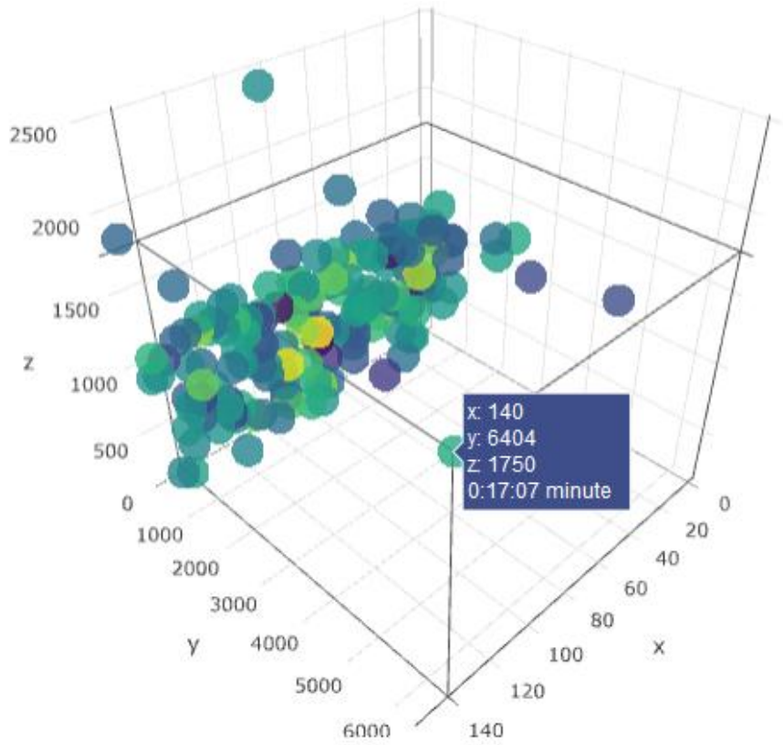
Практически не делают!

Иногда, если объектов мало и можно интерактивно вращать

3D-визуализации

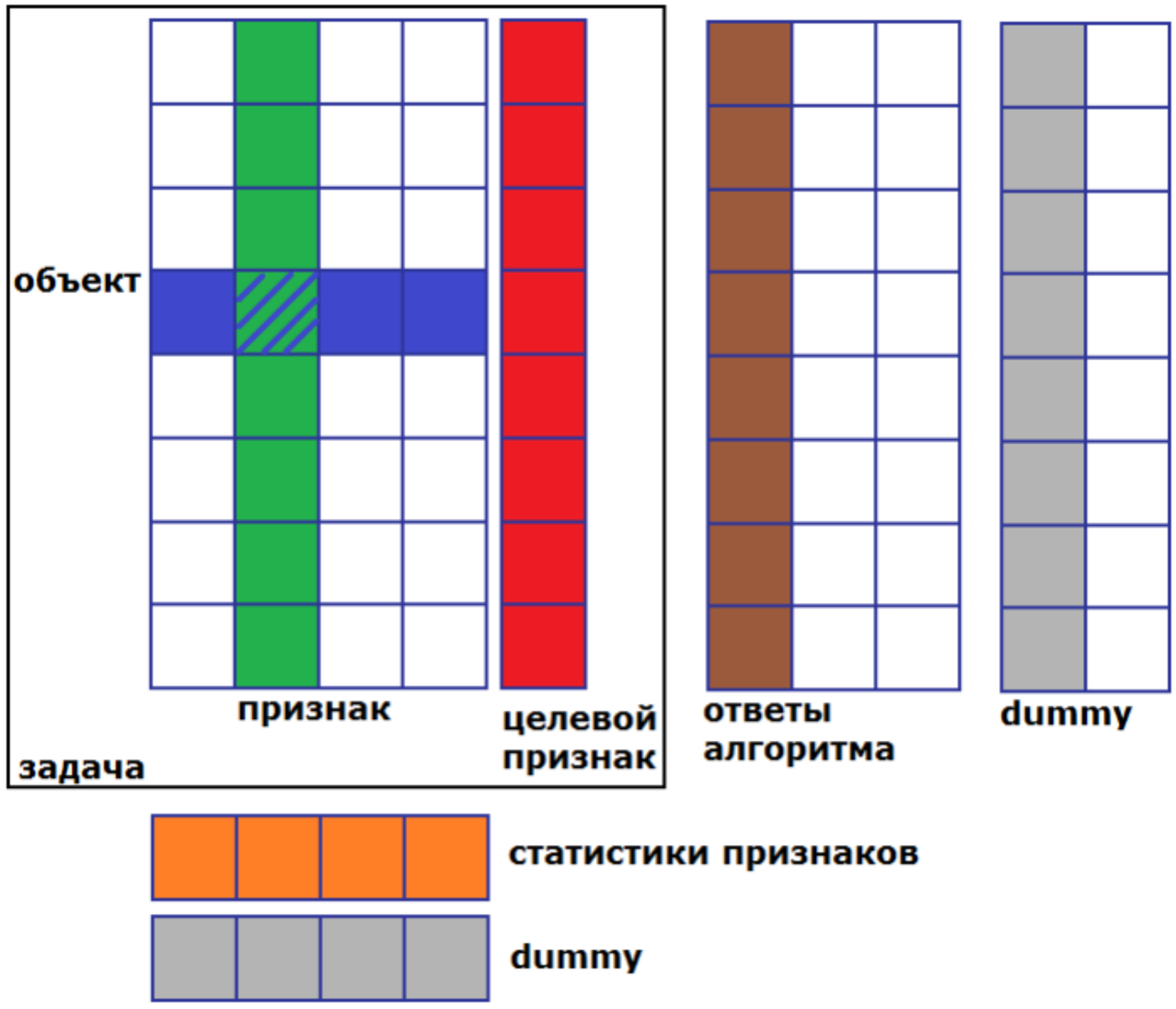
Пример, что не всегда размеры наглядны…





https://www.kaggle.com/ujjwal9/eda-passnyc
https://www.kaggle.com/saduman/eda-and-data-visualization-with-plotly

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.)

Пример визуализации служебных признаков

Сделайте график «id – время»:

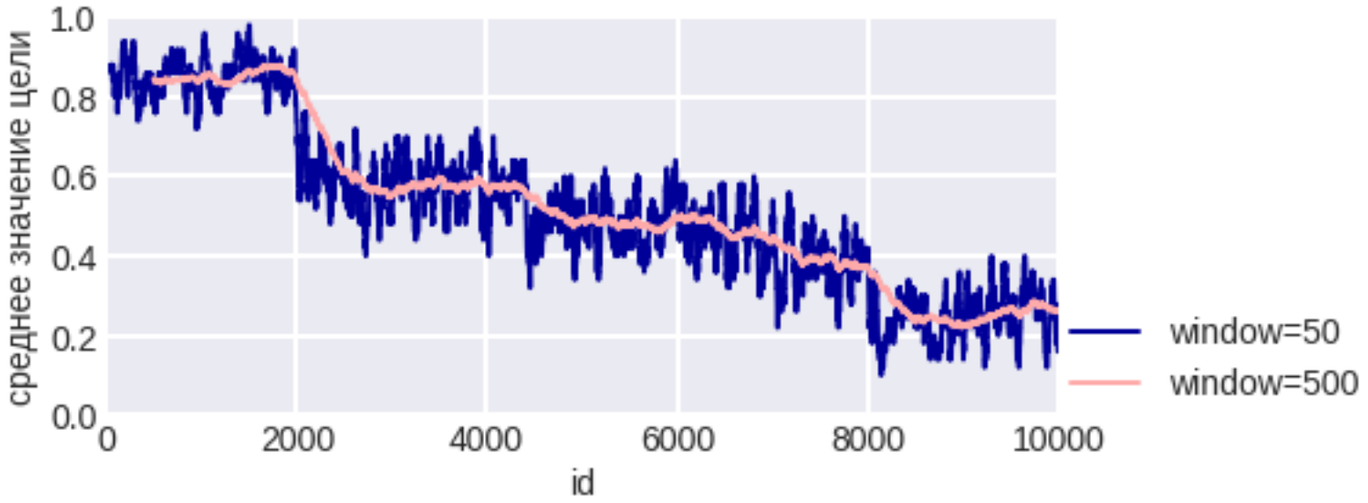
- простая проверка на монотонность
- видны «подозрительные периоды»

Случай из жизни: время – номер объекта

Видна двойная нумерация, периоды непоявления объектов

При раскраске по другим признакам видно больше!

Пример визуализации служебных признаков



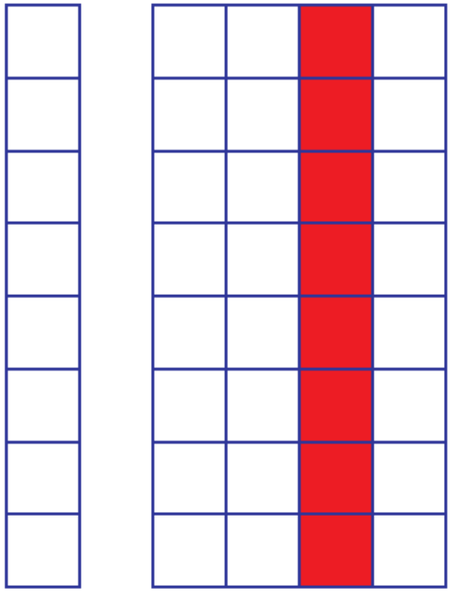
Как меняется цель со временем

```
plt.plot(df.y.rolling(50).mean())
plt.plot(df.y.rolling(500).mean())
```

видно перемешивались ли признаки…

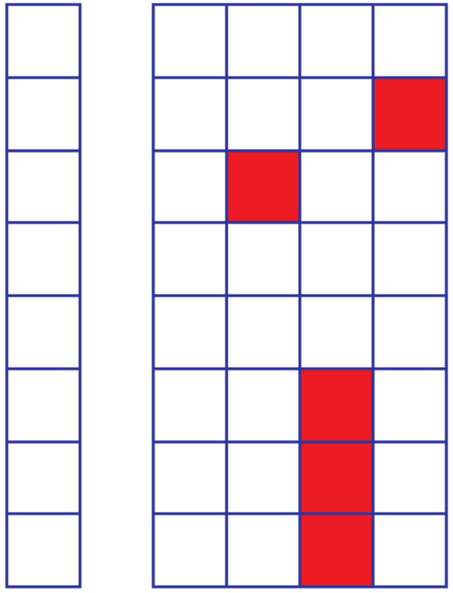
Что есть в данных:

- шумовые признаки



удалить

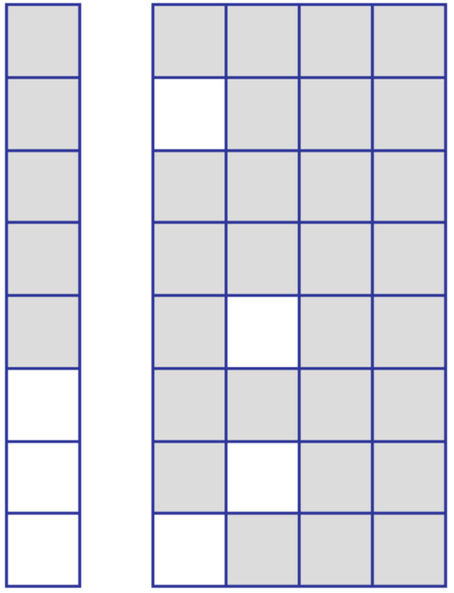
- шумовые значения



причины:

«ошибки из-за невнимательности»,
«особые режимы»

-пропуски:



причины:

«нет значения»,
«не знаем значения»

метод:

+ служебные признаки / данные

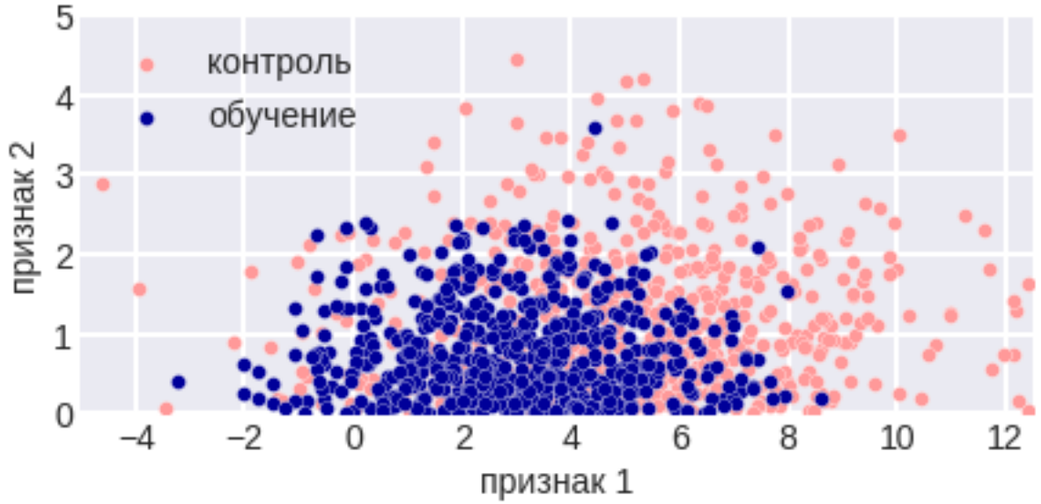
Ещё приёмы в EDA...

Проверка соответствия «train-test»

Что надо проверить найдя закономерность?

Что «контроль» ложится на обучение!

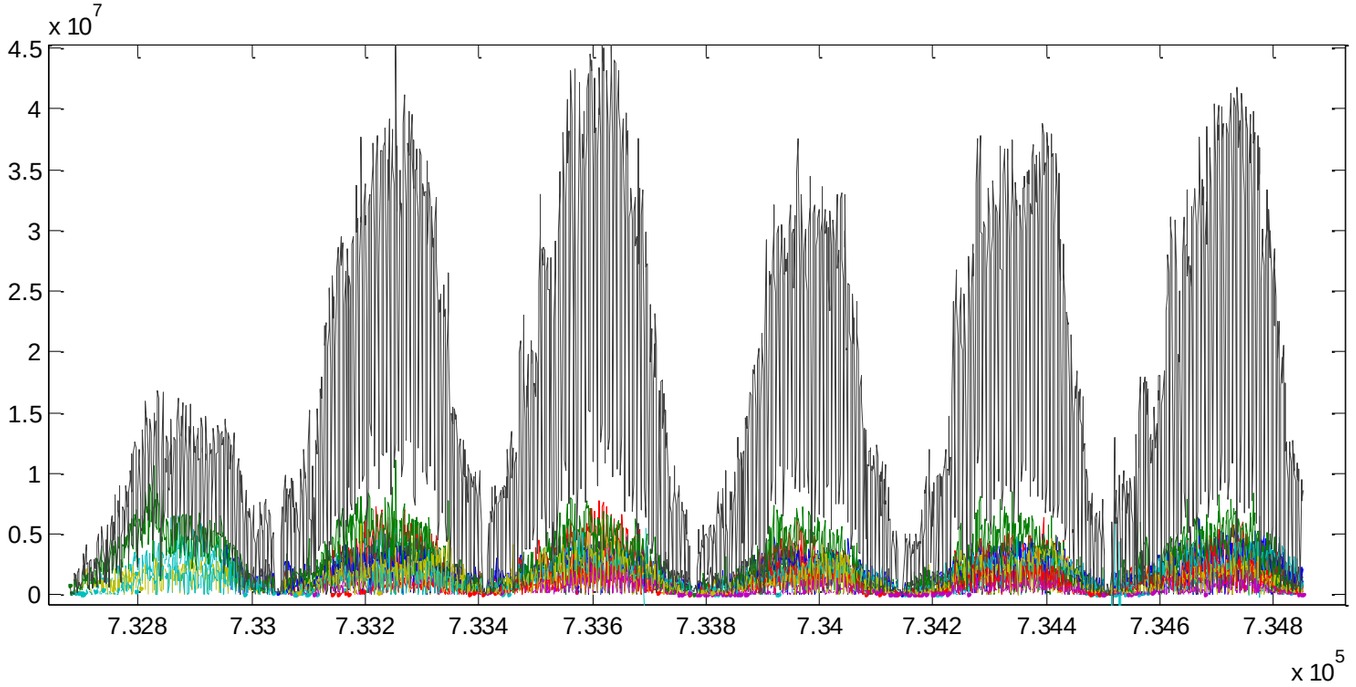
На практике нет гарантий одинаковости распределений, даже если это гарантирует заказчик.



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

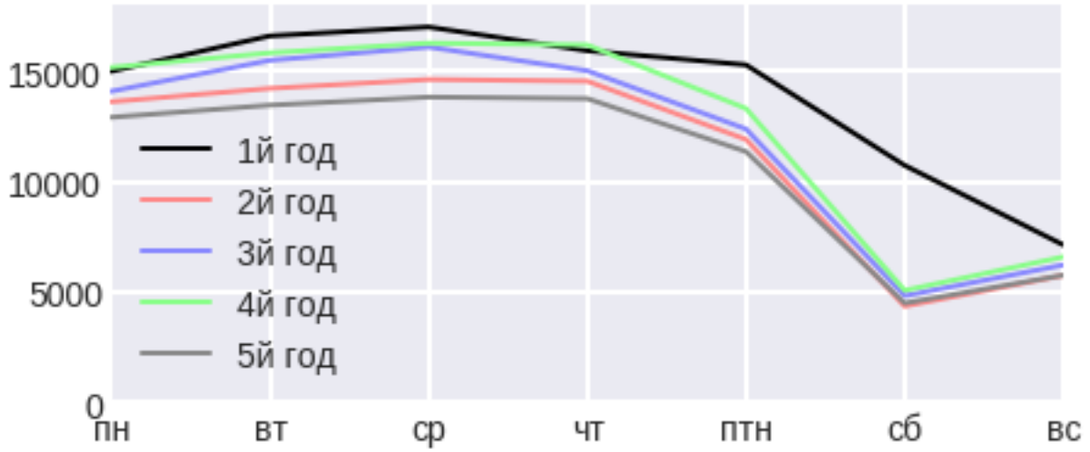
Агрегация (по дням недели)

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

Удивительно, но при визуализации:

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

Если этого нет:

- ищем ошибку

Итог

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...