Прикладные задачи анализа данных

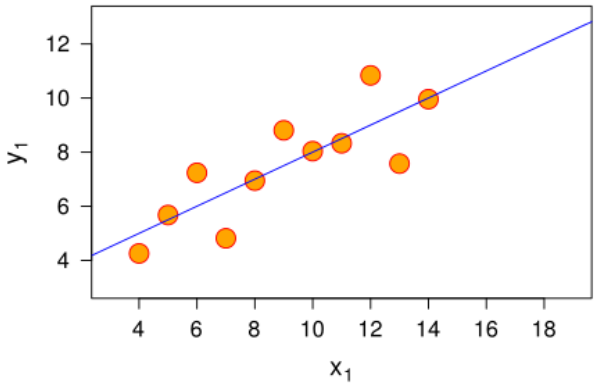
ИСКУССТВО ВИЗУАЛИЗАЦИИ

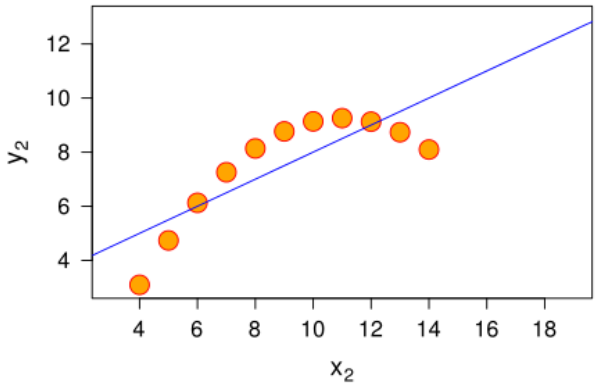
Дьяконов А.Г.

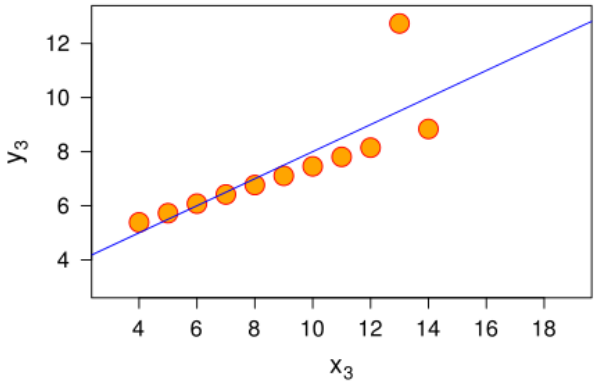
Московский государственный университет имени М.В. Ломоносова (Москва, Россия)

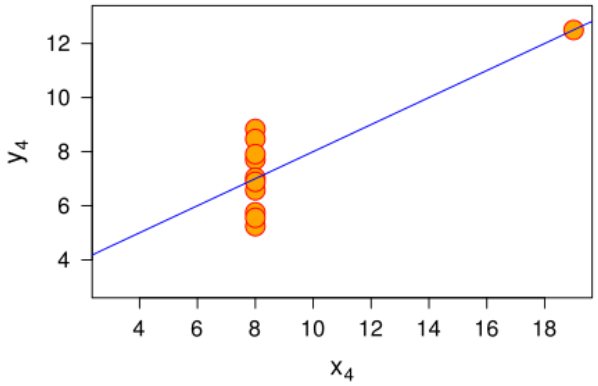
Зачем смотреть на данные?

Наборы данных имеют идентичные статистические характеристики, но их графики существенно различаются.







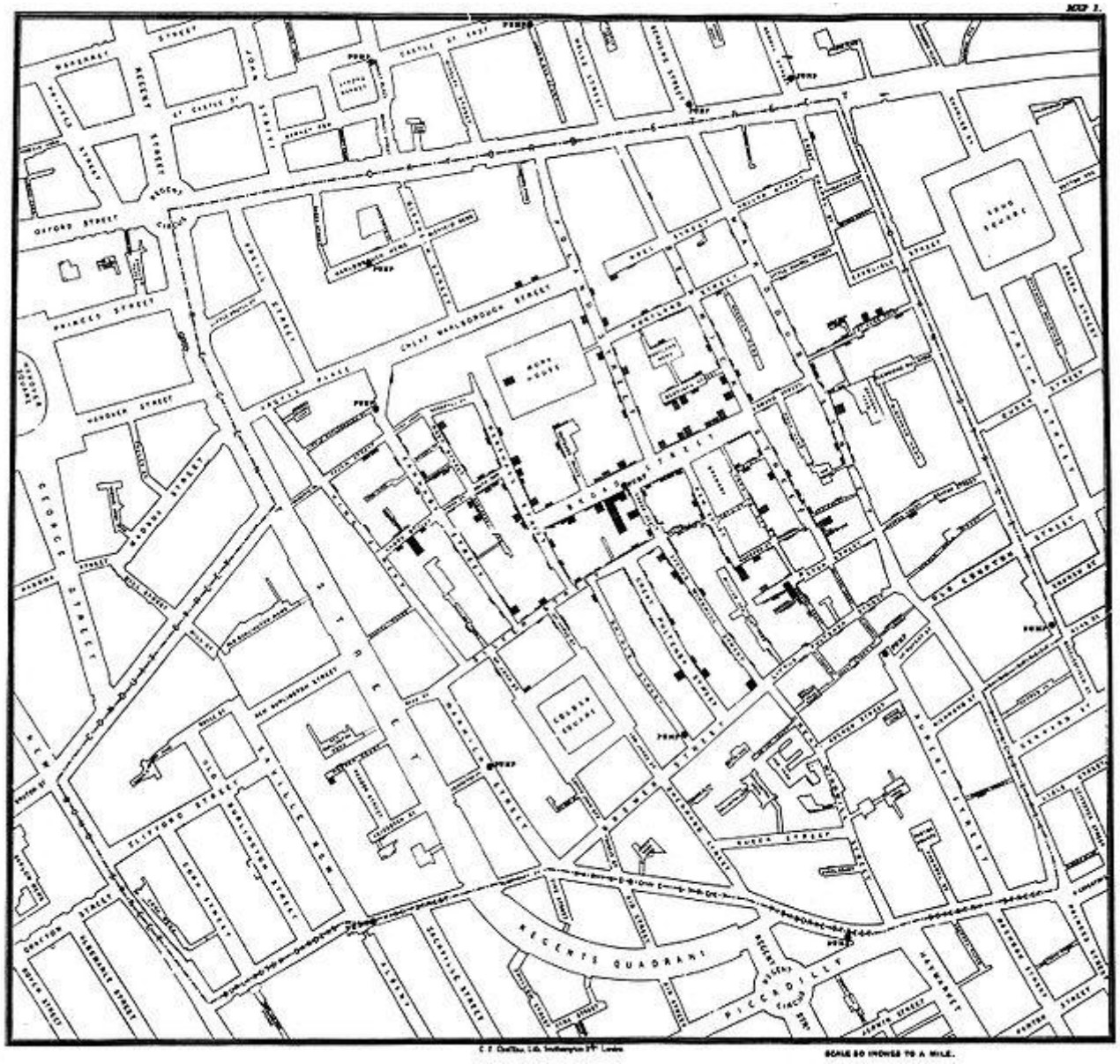


Характеристика	Значение
Среднее значение переменной X	9.0
Дисперсия переменной X	10,0
Среднее значение переменной Y	7,5
Дисперсия переменной Y	3,75
Корреляция между переменными	0,816
Прямая линейной регрессии	Y=3+X/2

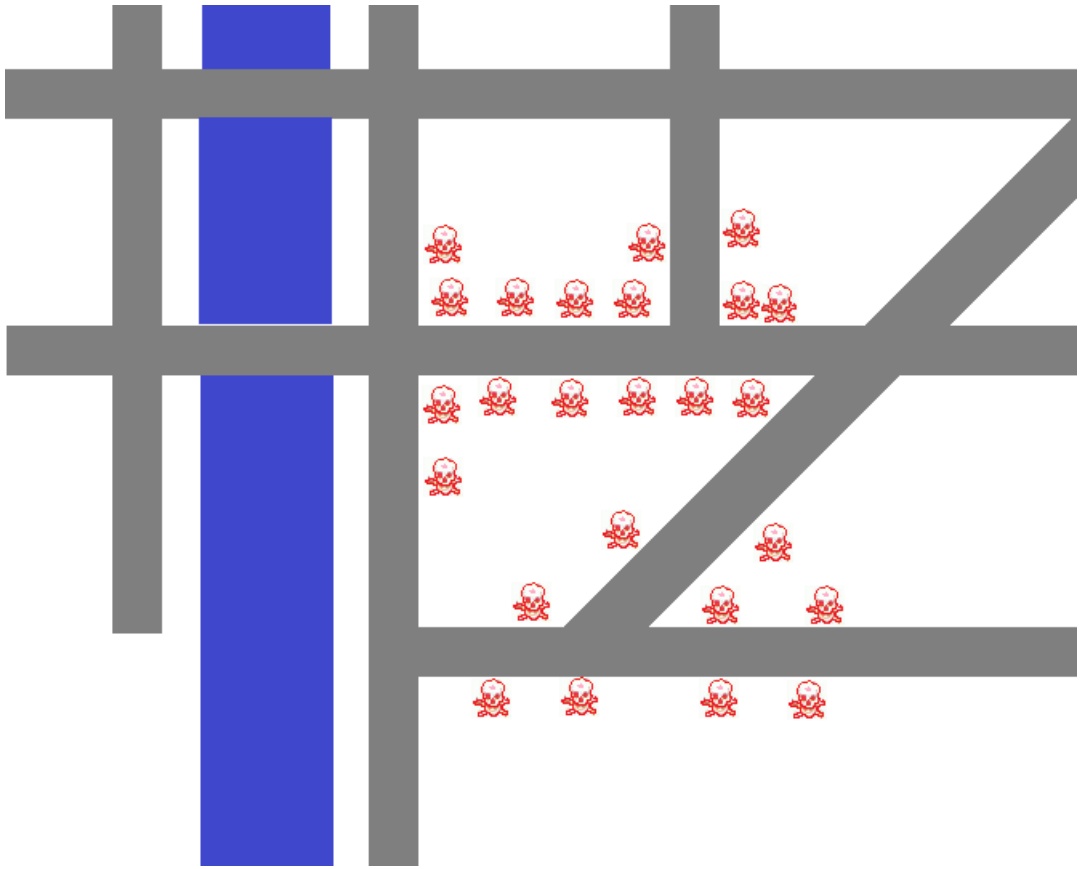
F.J. Anscombe Graphs in Statistical Analysis // American Statistician, 27 (February 1973), 17-21.

Вспышка холеры на Брод-стрит в 1854 году

См. Википедию



Статистика заболевания холерой



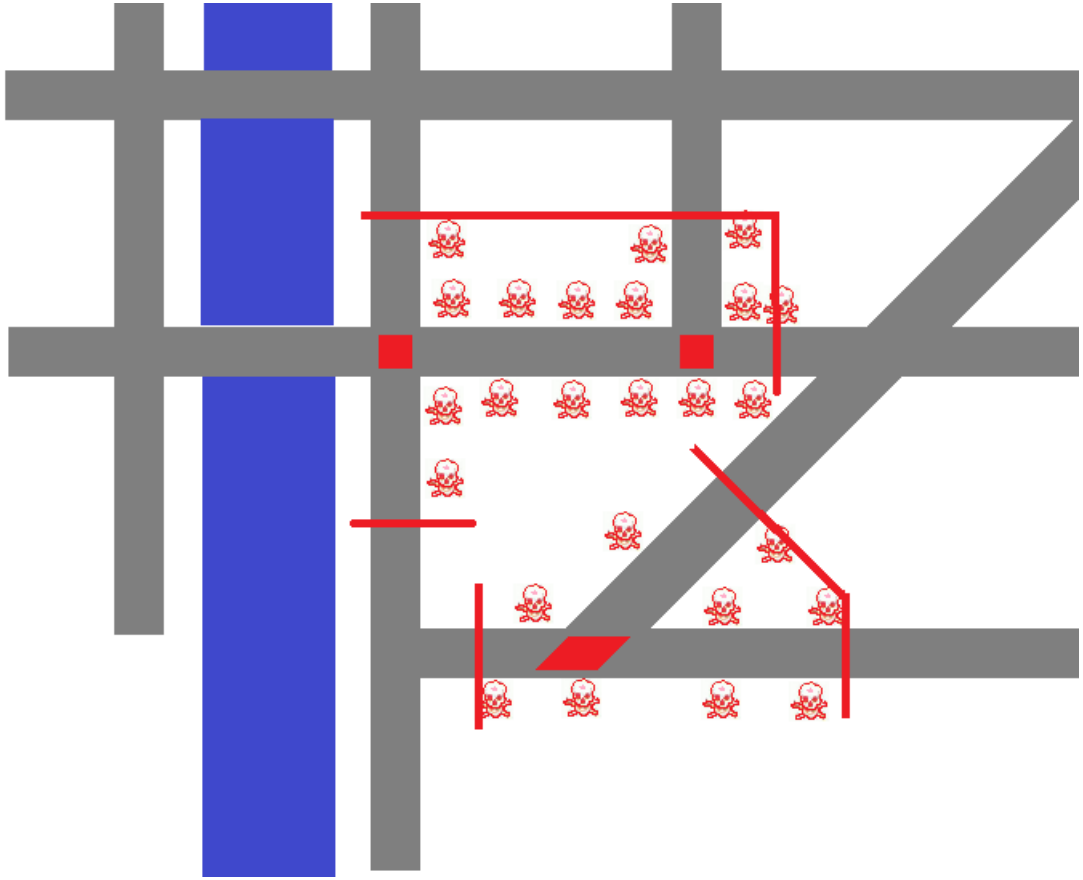
Всего умерло 616 человек!

Причина?!

Кто такой Джон Сноу?

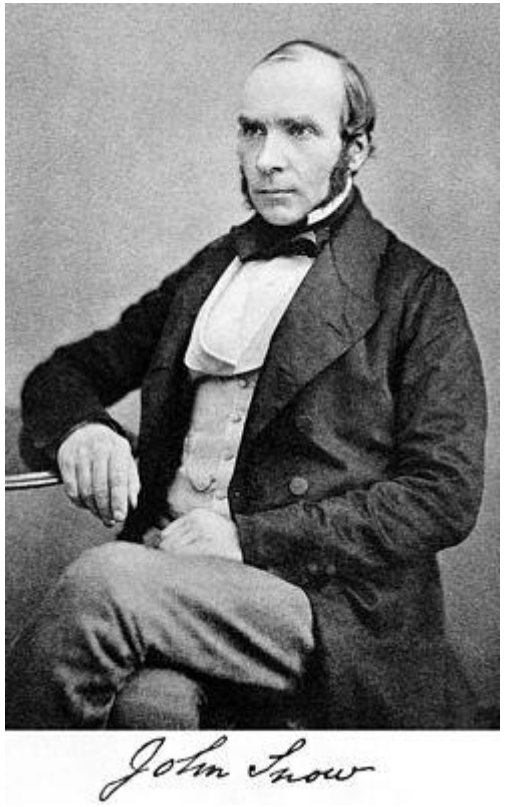
Центры эпидемии – колодцы!

Диаграмма Вороного была видна на карте…



Нечистоты сливались в Темзу,
в результате была заражена
местная система водоснабжения.

Джон Сноу



(15.03.1813 — 16.06.1858)
британский врач, один из пионеров массового внедрения анестезии и медицинской гигиены

Что это за данные?

Что это за данные?

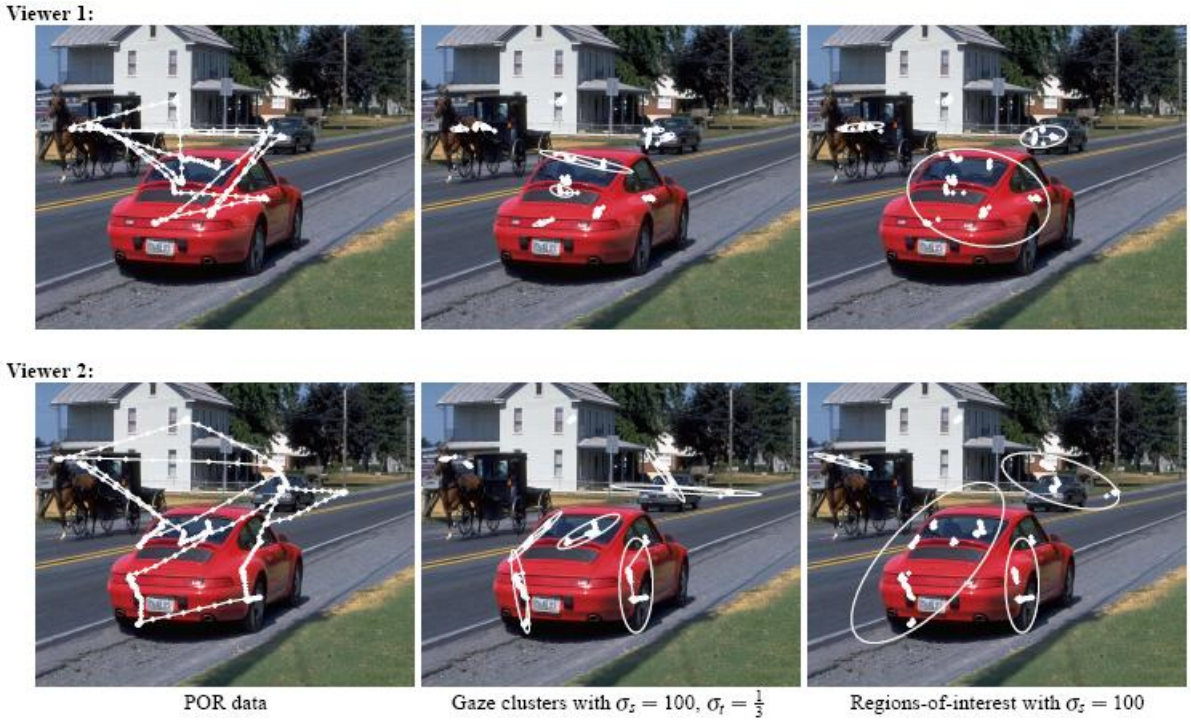
Ответ

Плотность внимания на Интернет-странице

Измеряется по числу и продолжительности фиксаций

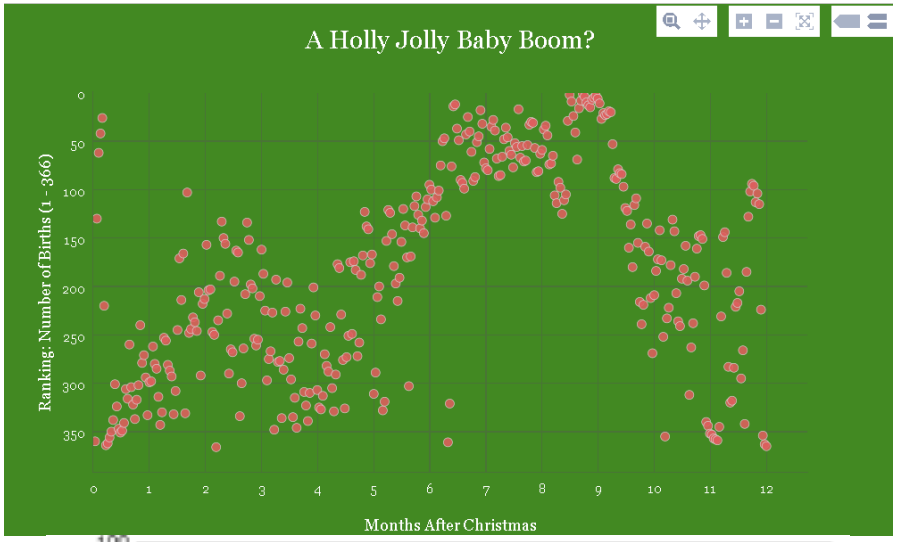
L. A. Granka, H. A. Hembrooke, G. Gay, M. K. Feusner Correlates of Visual Salience and Disconnect: An Eye-tracking Evaluation

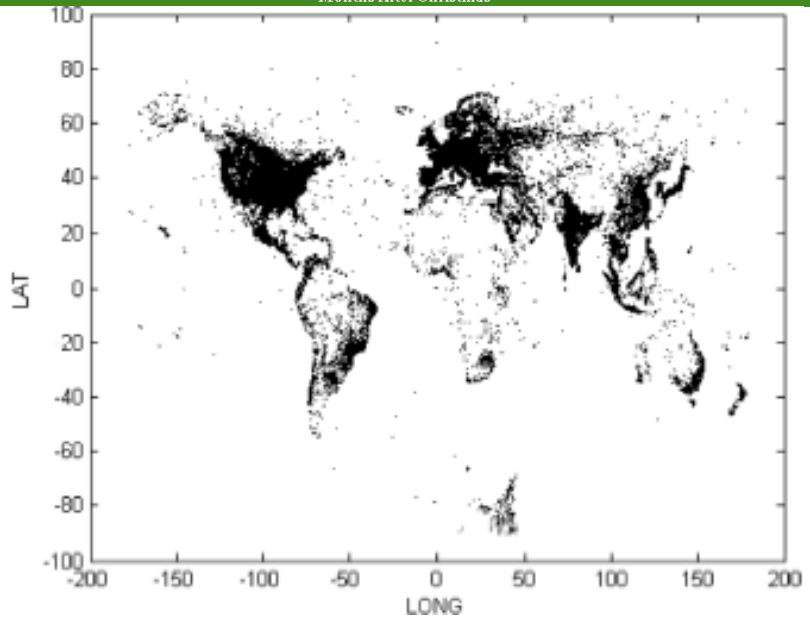
A. Santella D. DeCarlo Robust Clustering of Eye Movement Recordings for Quantification of Visual Interest



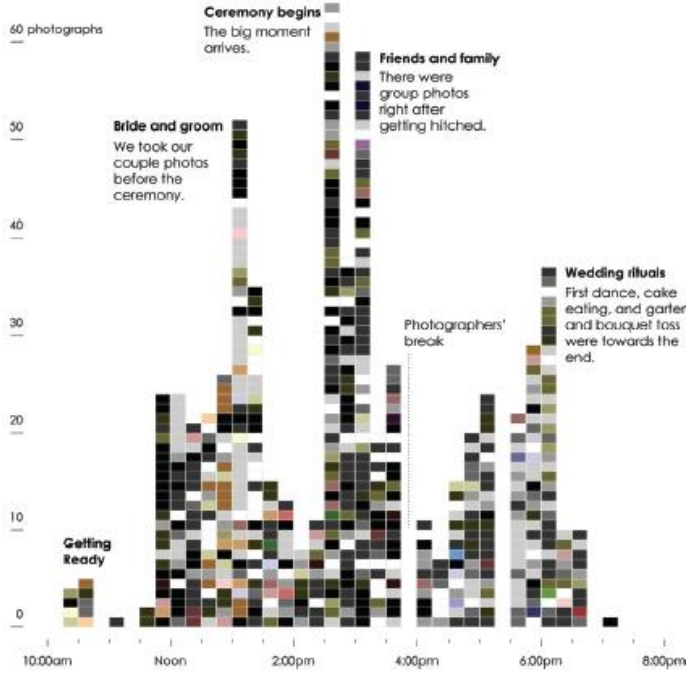
Домашнее задание

Найти интересные нетривиальные визуализации

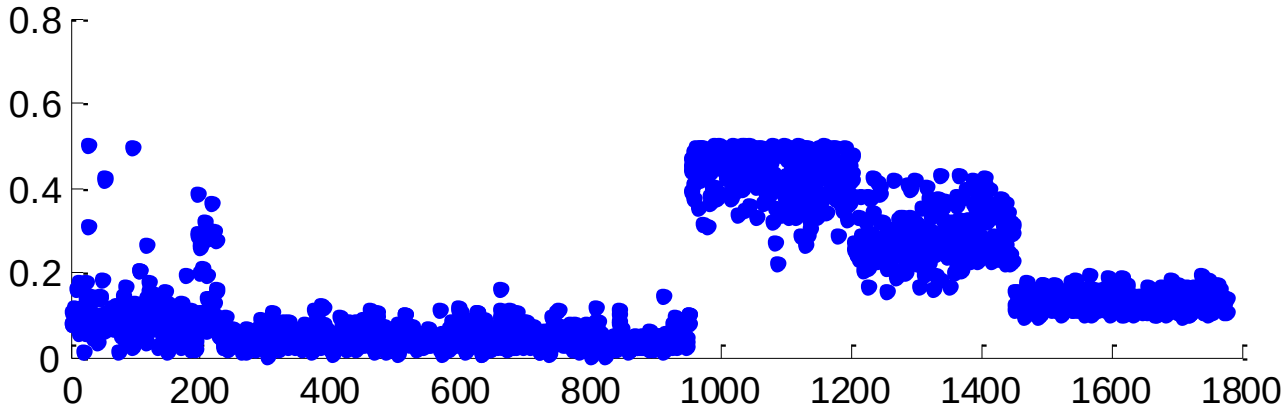


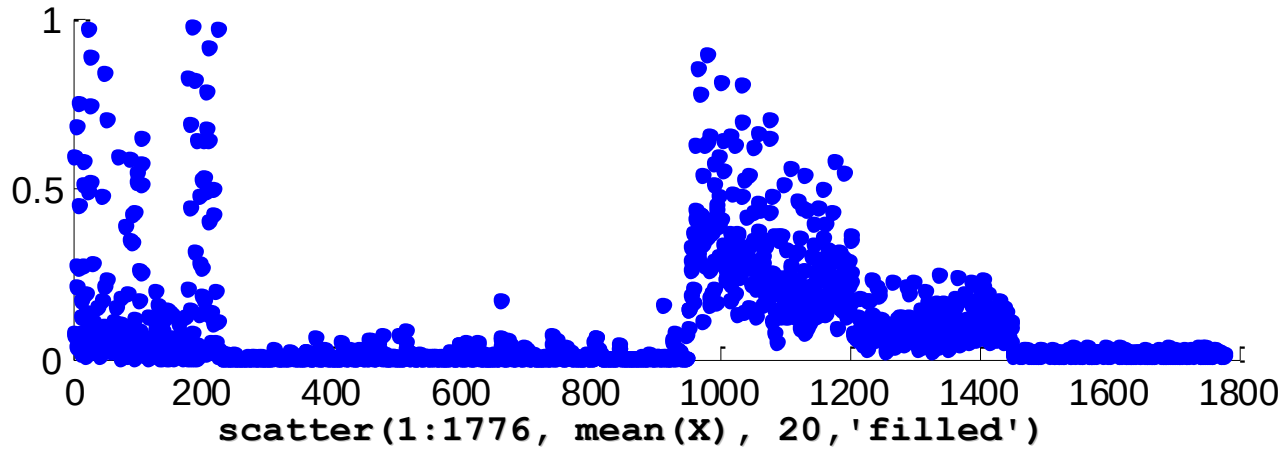


GPS-координаты



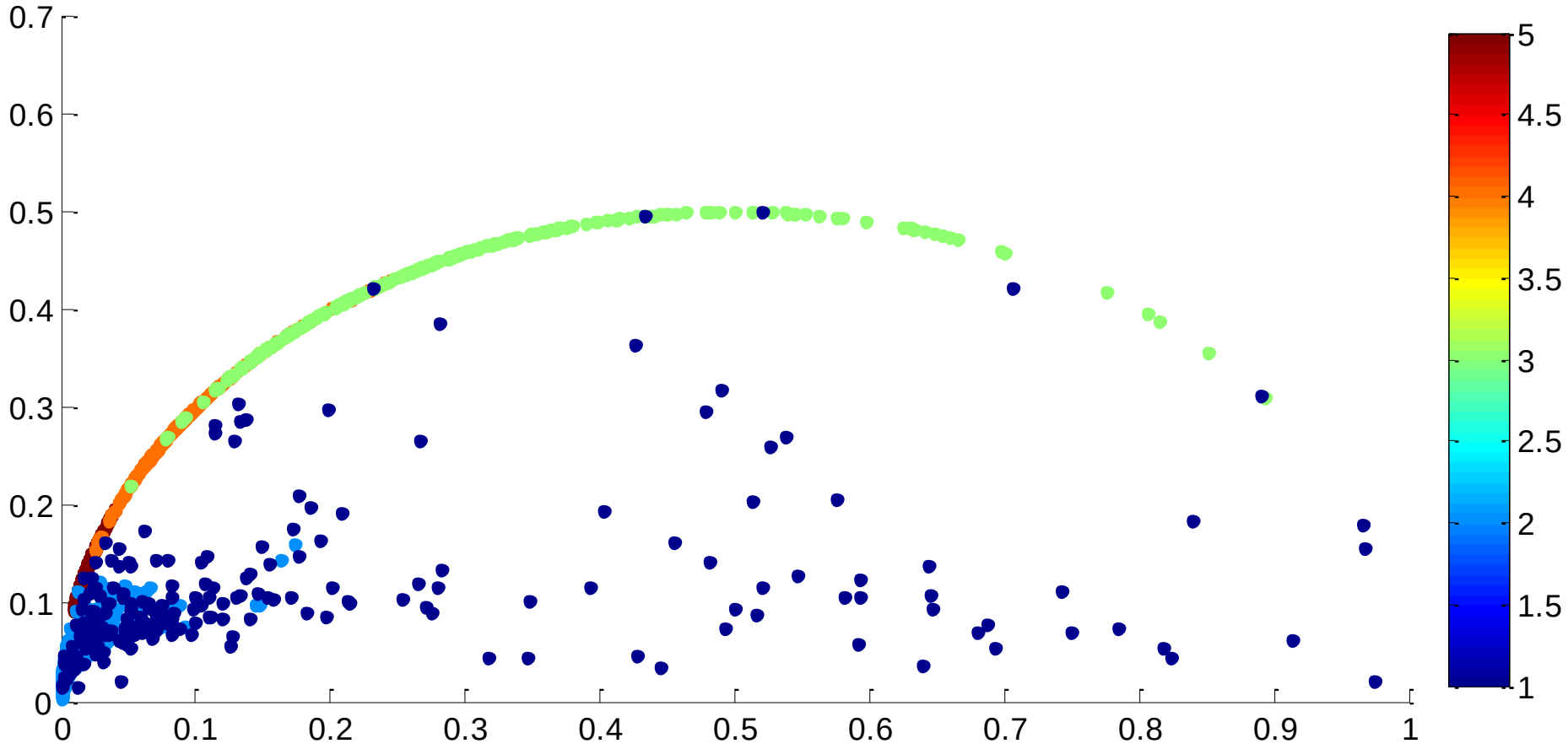
ЗАДАЧА BIOLOGICAL RESPONSE





Чётко видны группы

Фантастика: дугообразная зависимость у трёх групп признаков!




```
Col = ismember(1:1776,225:951)+2*ismember(1:1776,952:1203)+
3*ismember(1:1776,1204:1449)+4*ismember(1:1776,1450:1776);
scatter(mean(X), std(X), 20, Col+1,'filled')
```

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]

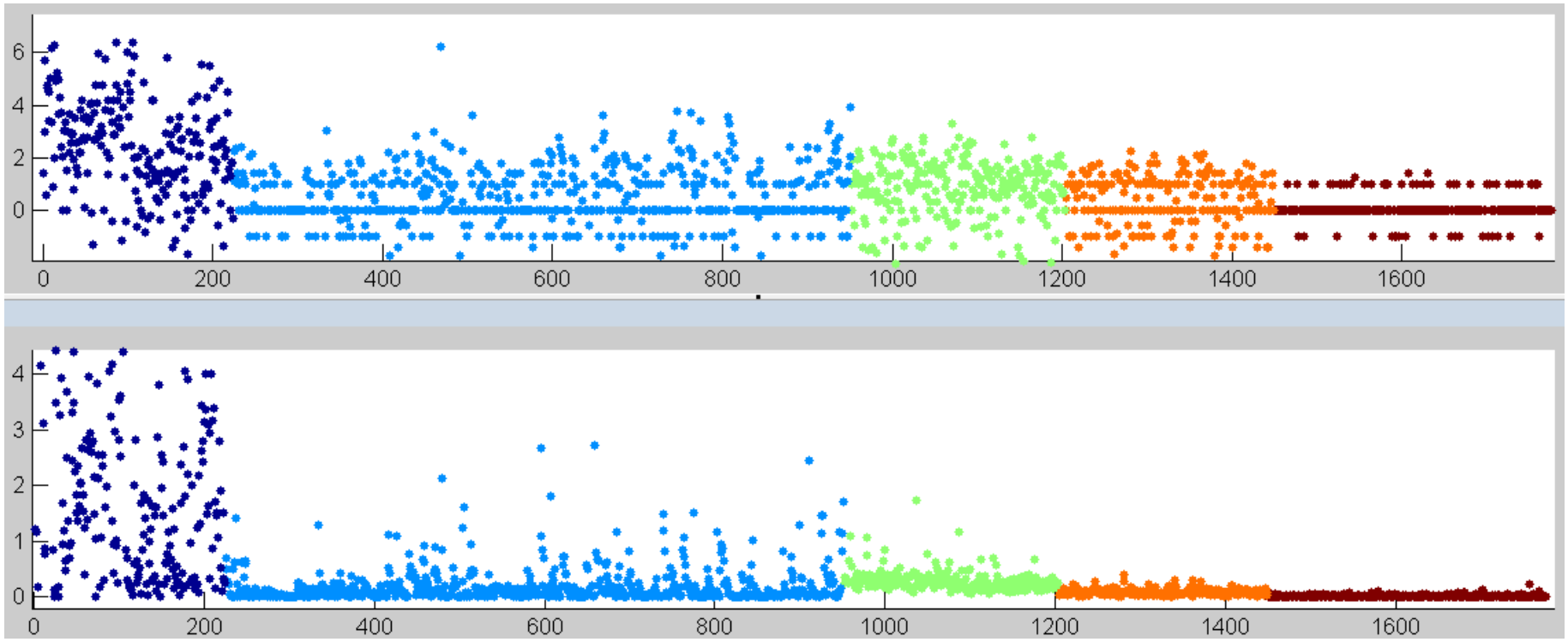
$$\text{mean}\{x_i\}_{i=1}^{n} = \frac{1}{n}\sum_{l=1}^{n} x_i \equiv p$$

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

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

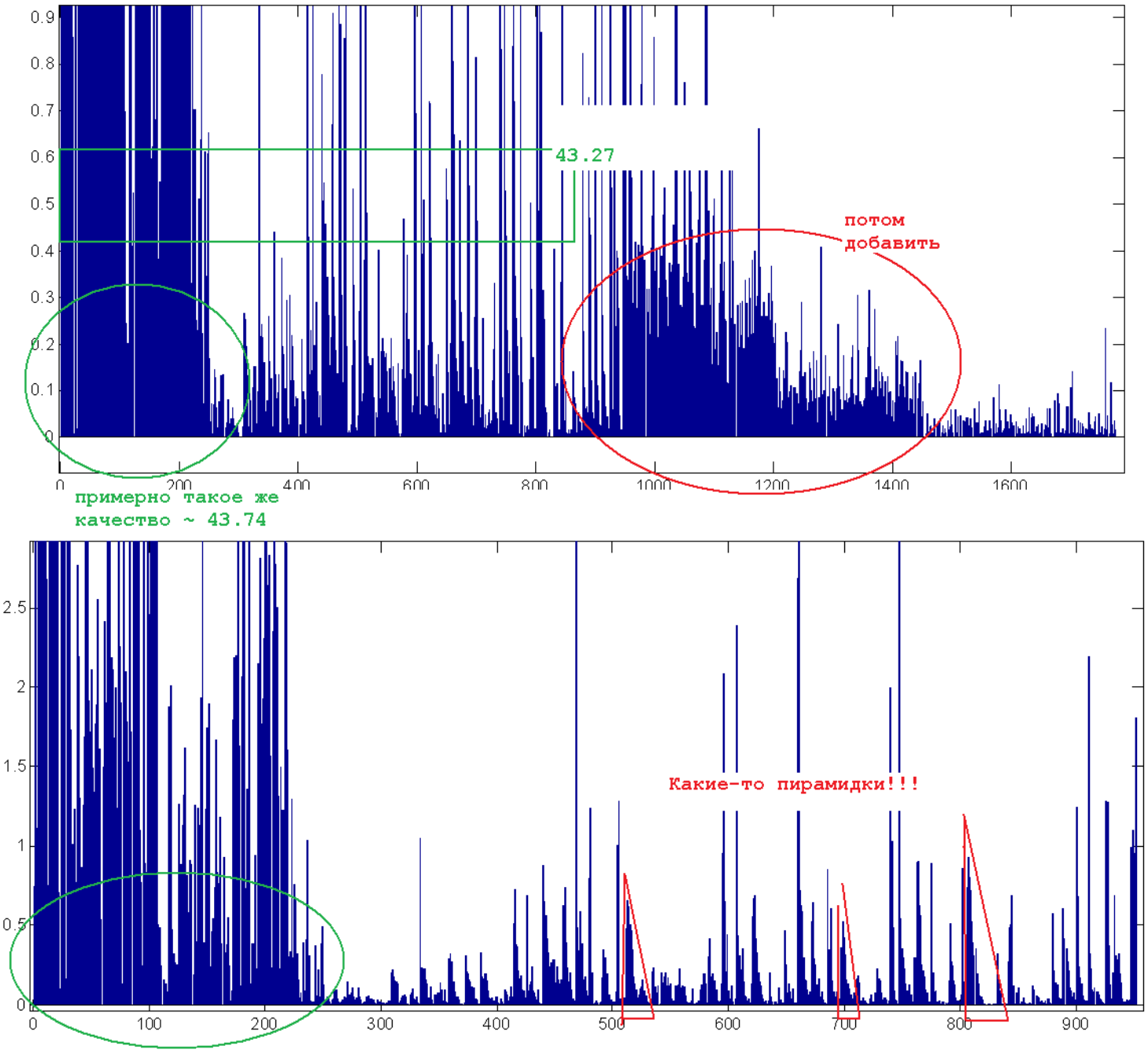
$$= \sqrt{\frac{1-2p}{n}\sum_{i=1}^{n} x_i + p^2} = \sqrt{(1-2p)p + p^2} = \sqrt{p - p^2} = \sqrt{p(1-p)}$$

Важности признаков с точки зрения RF.



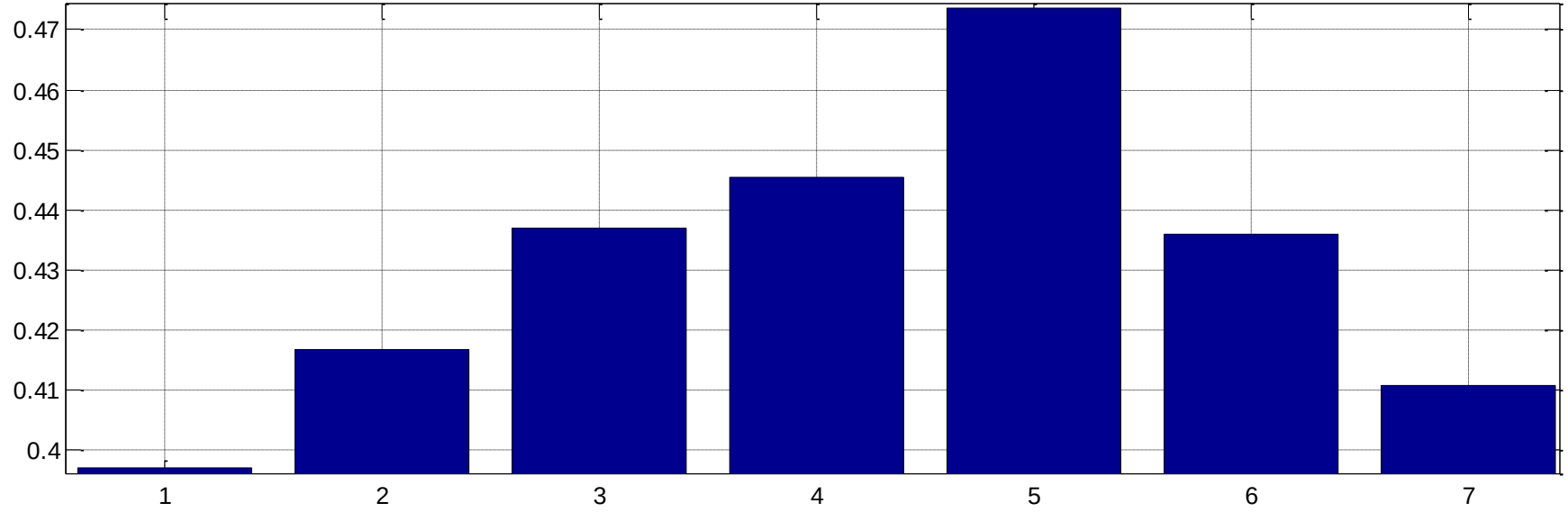

```
scatter(1:1776,  importance1(:,2),20,Col+1,'filled')
scatter(1:1776,  importance1(:,3),20,Col+1,'filled')
```

Потом: целые группы признаков можно удалять без существенной потери качества


43.27
потом
добавить
примерно такое же
качество ~ 43.74
Какие-то пирамидки!!!

Аналогично: исследование сложности «классификации» объектов

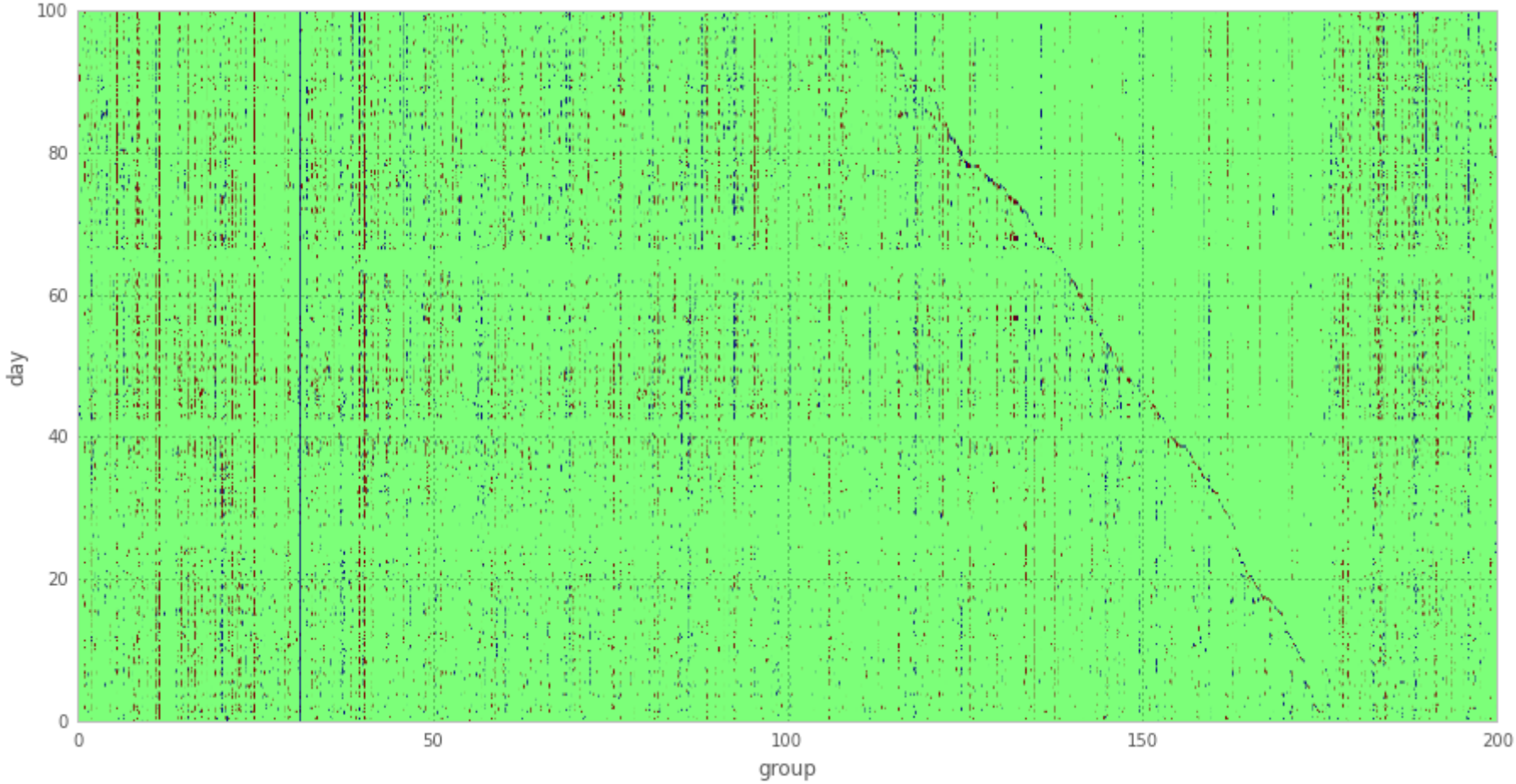
Исследование частей выборки(фолдов)



Сложность 500-задачек с точки зрения моего RF-а

[0.3971 0.4168 0.4370 0.4455 0.4736 0.4359 0.4109]

Визуализация данных (RedHat)



по горизонтали – разные группы,

по вертикали – дни (подряд),

салатовый цвет – нет взаимодействия,

красный / синий – класс 1 / 0

Что за подозрительная полоса?

Визуализация данных (RedHat)

Группы упорядочены так:

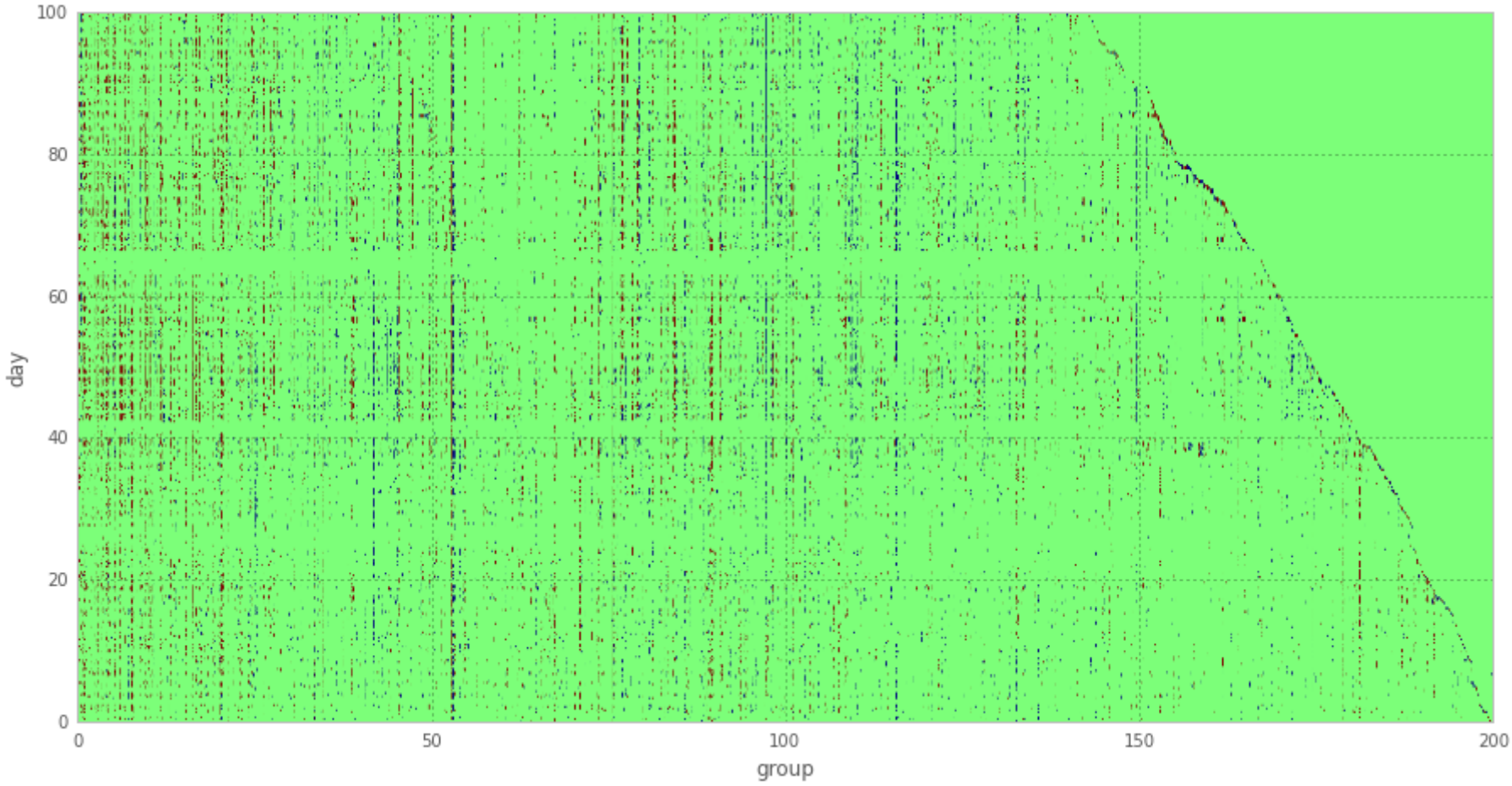
```
group_date2.columns[:10]

'group 1000', 'group 10006', 'group 1001', 'group 1002', 'group
10021', 'group 10025', 'group 10032', 'group 10036', 'group 1004',
```

это лексикографический порядок!

Теперь сделаем в обычном порядке…

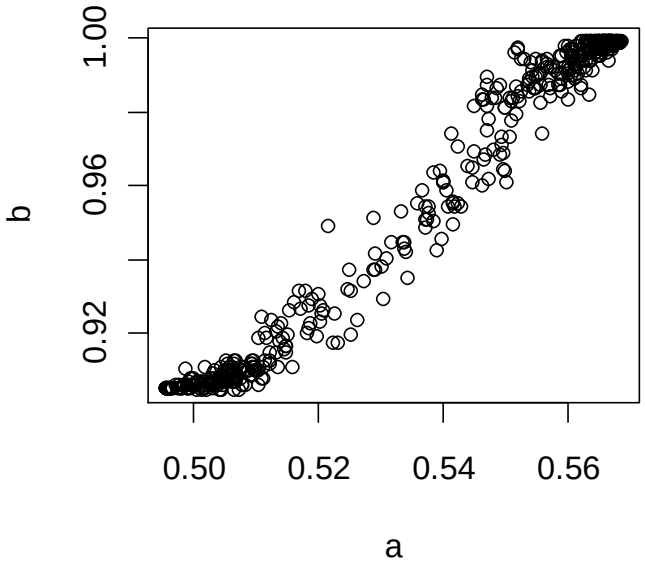
```
data_train.group_1 = data_train.group_1.map(lambda x: int(x[6:]))
```

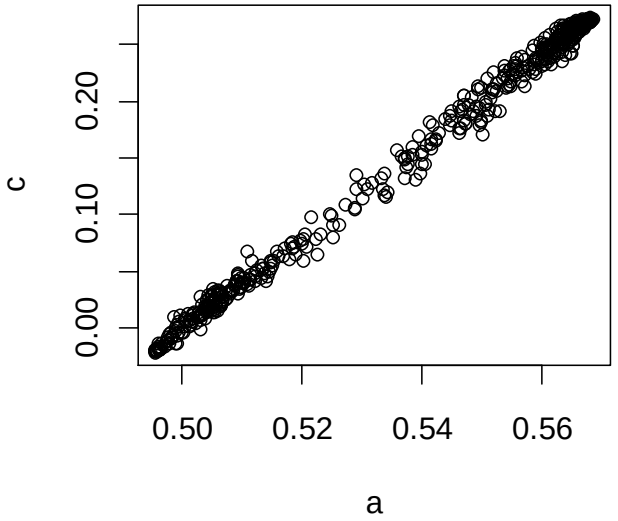
Визуализация данных (RedHat)



теперь понятнее… группы, видимо, идут в порядке появления последние – которые добавлялись в дни сбора выборки

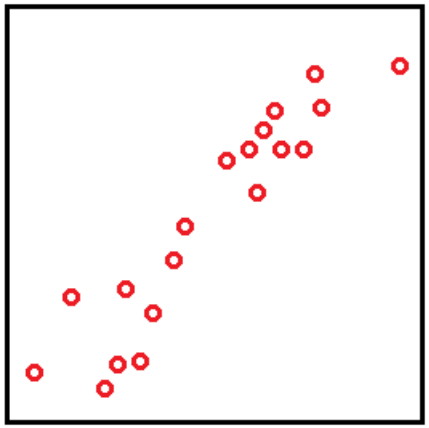
Ответы алгоритмов – целевые значения





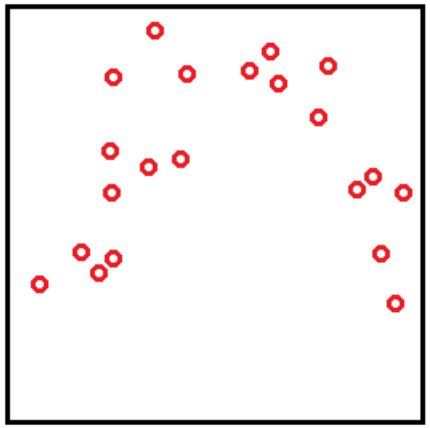

```
LogPer <- function(a, y)
                    {
a = pmin( pmax(a,0.0001), 0.9999);
mean(- y*log(a) - (1-y)*log(1-a))
                    }
```

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



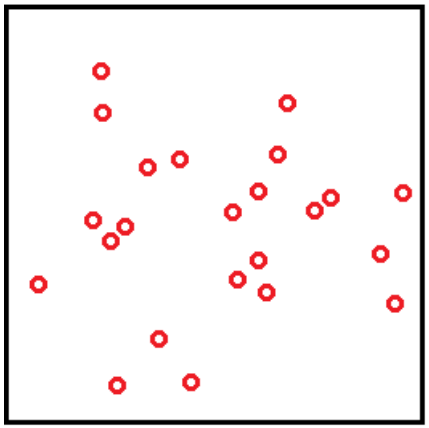
Корреляцию

сложно из-за масштаба



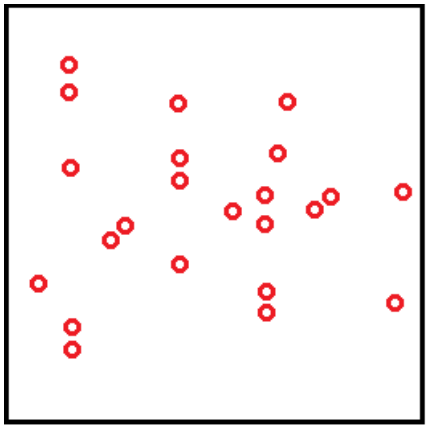
Зависимость

сложно из-за неравномерности, размазанности



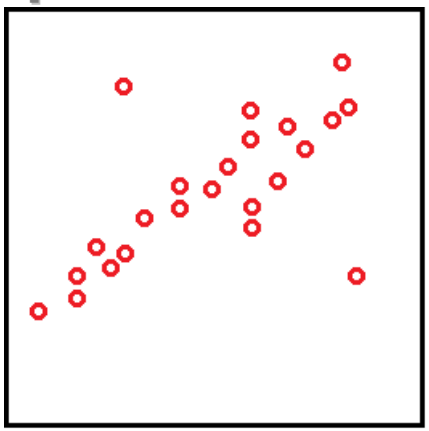
Независимость

Часто ложное видение



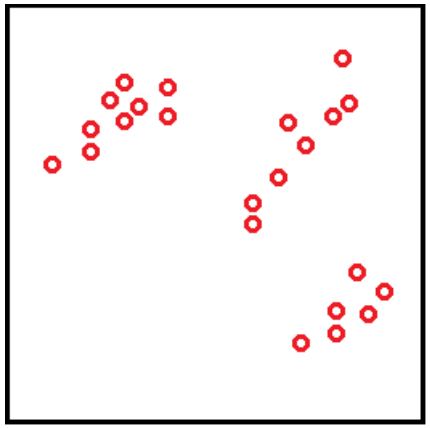
Уникальные значения

сложно из-за объёма, накладок



Выбросы

сложно из-за масштаба



Кластеры

сложно из-за масштаба, искусственности

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

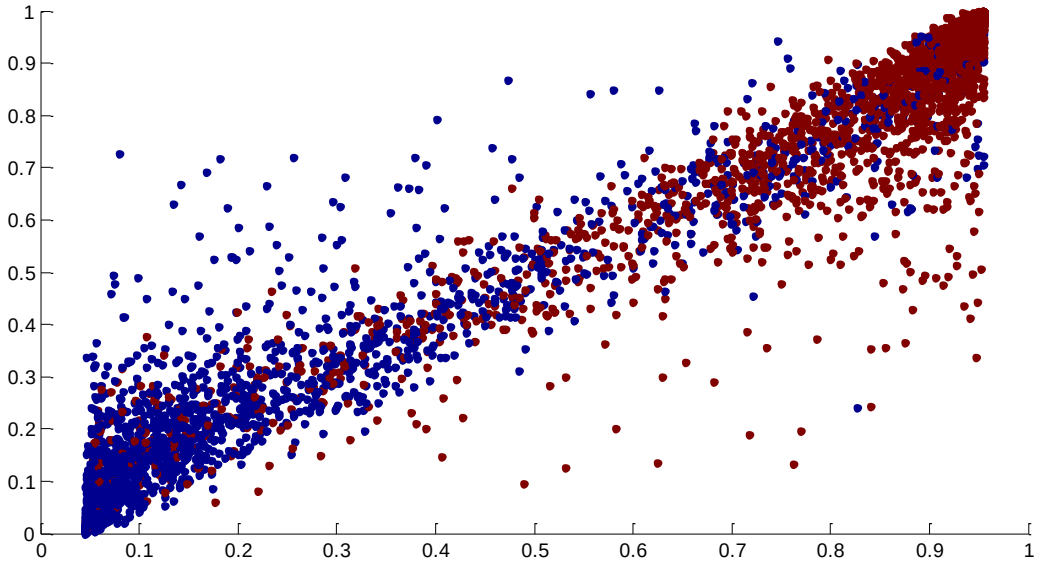
```
pd.crosstab(people.char_1, people.char_2)
```

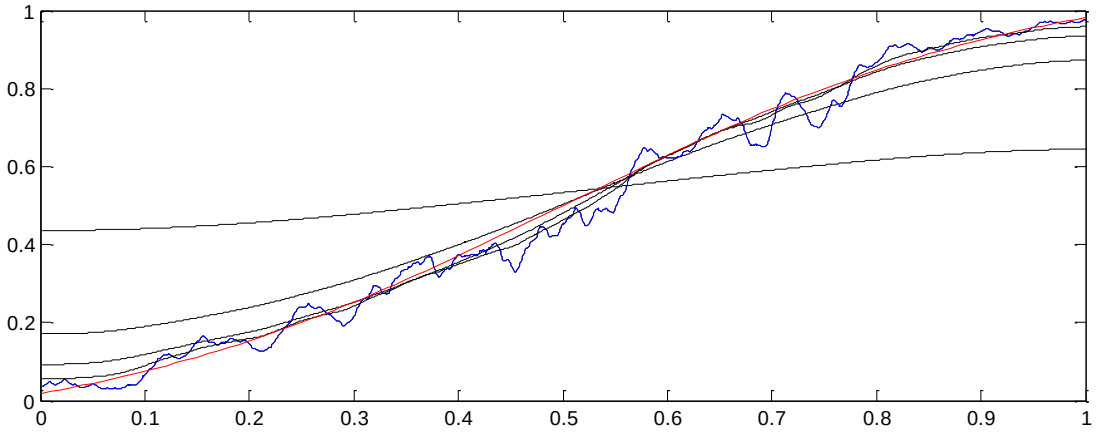
char_2	type 1	type 2	type 3
char_1			
type 1	15251	0	0
type 2	0	77314	96553

Как использовать это знание?

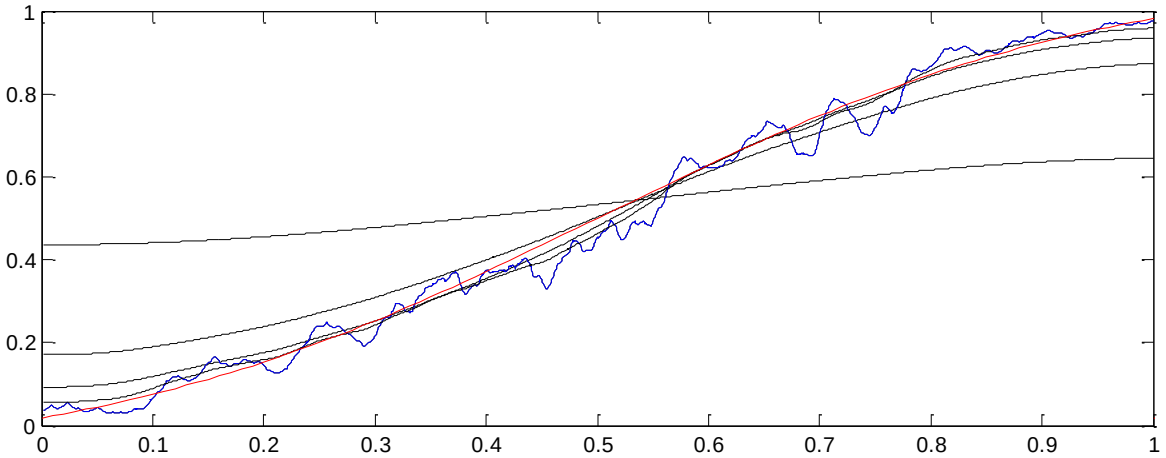
Реальная прикладная задача

«Biological Response»





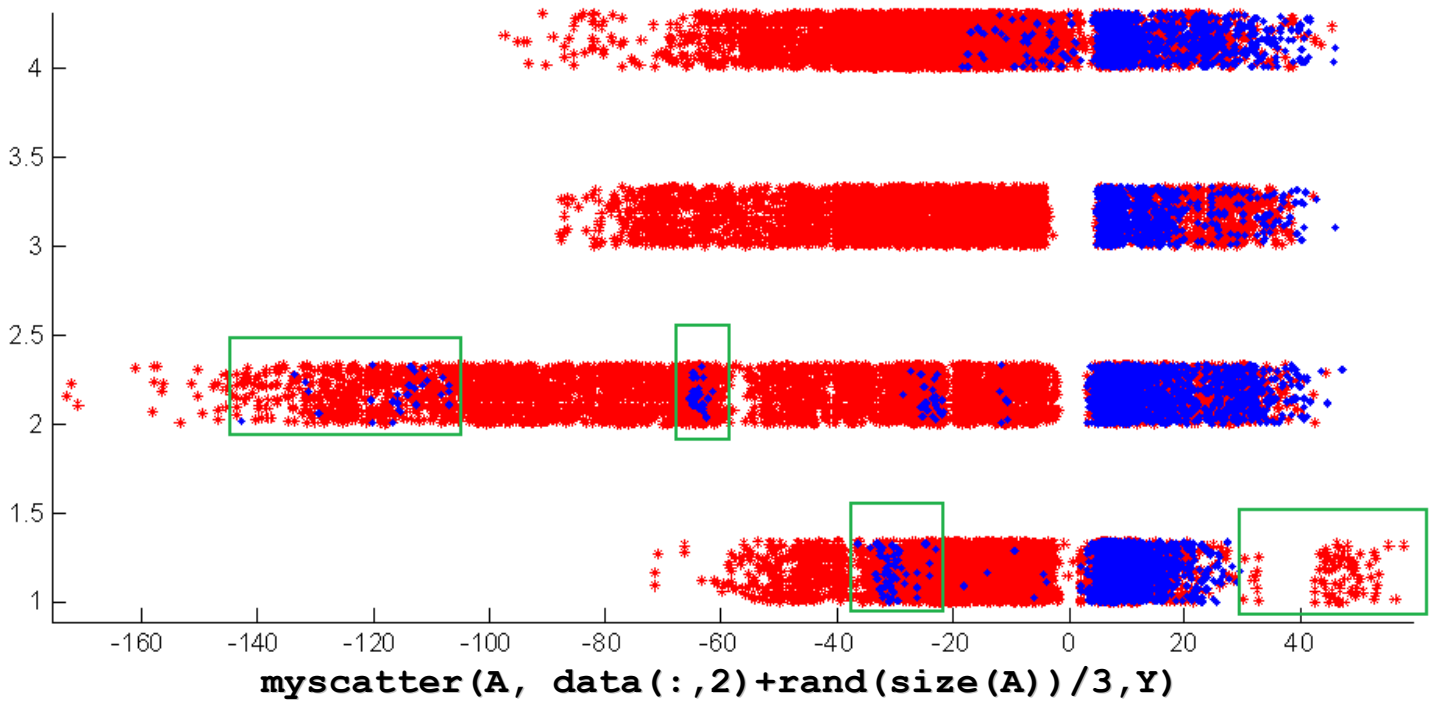
«Deformation of Random Forests»



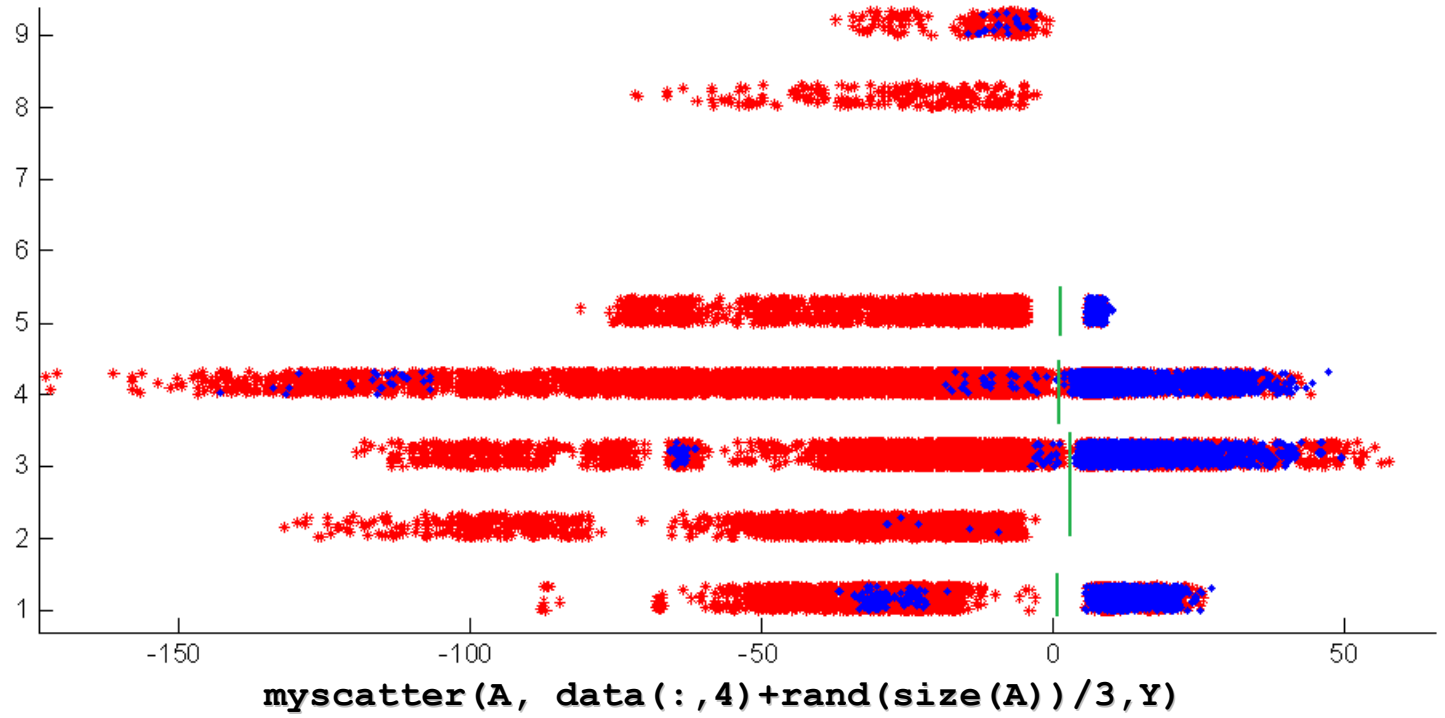
$$\beta\left(\frac{1}{1+e^{-\alpha(x-0.5)}}-0.5\right)+0.5$$

Задача «~Analytics»

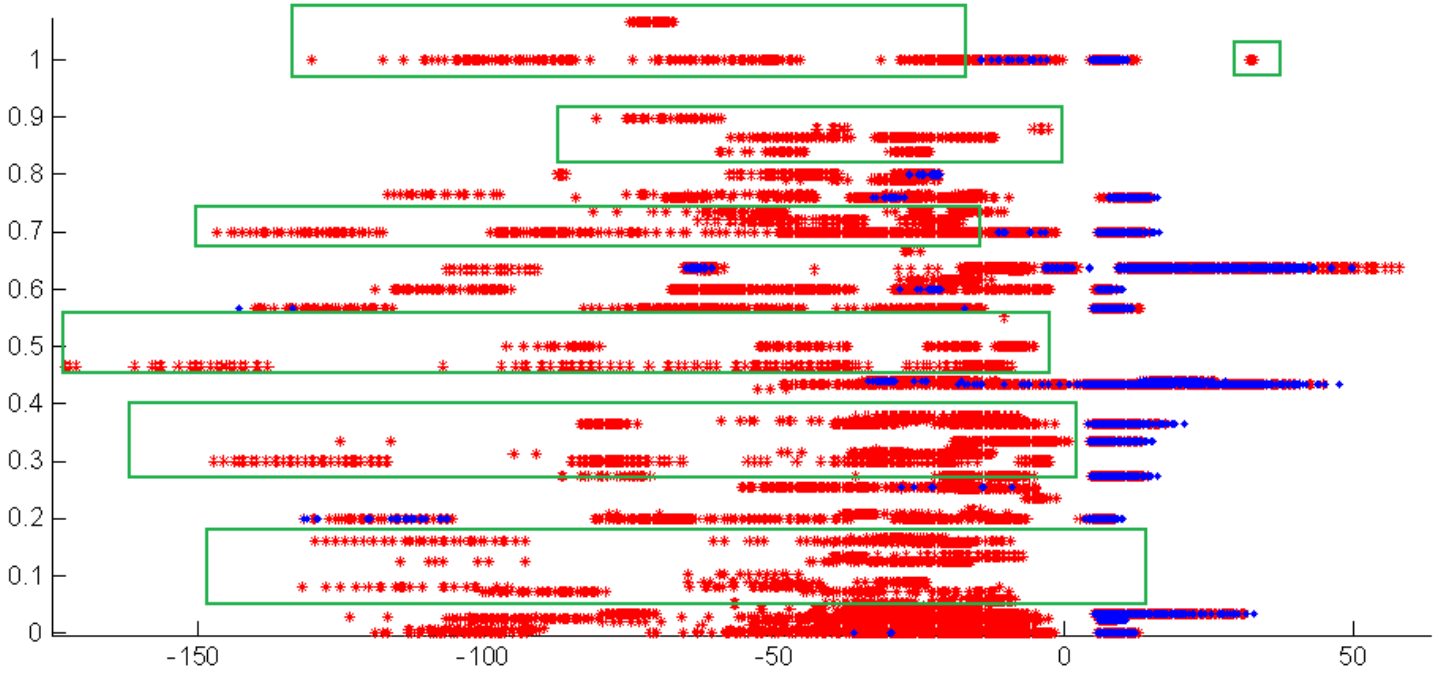
Ответы алгоритма – признак



Нахождение закономерностей – 1



Нахождение закономерностей – 2



Что надо проверить найдя закономерность?

Что надо проверить найдя закономерность?

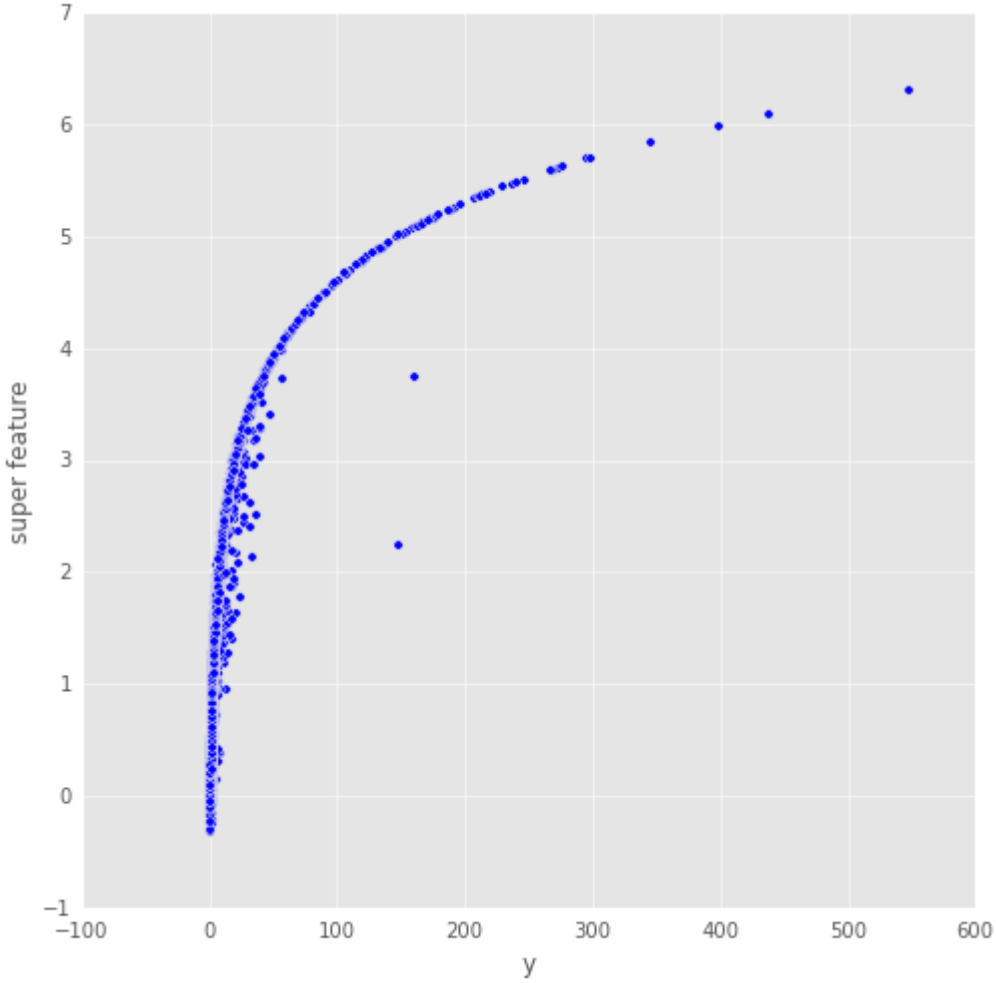
Что «контроль» ложится на обучение!

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?



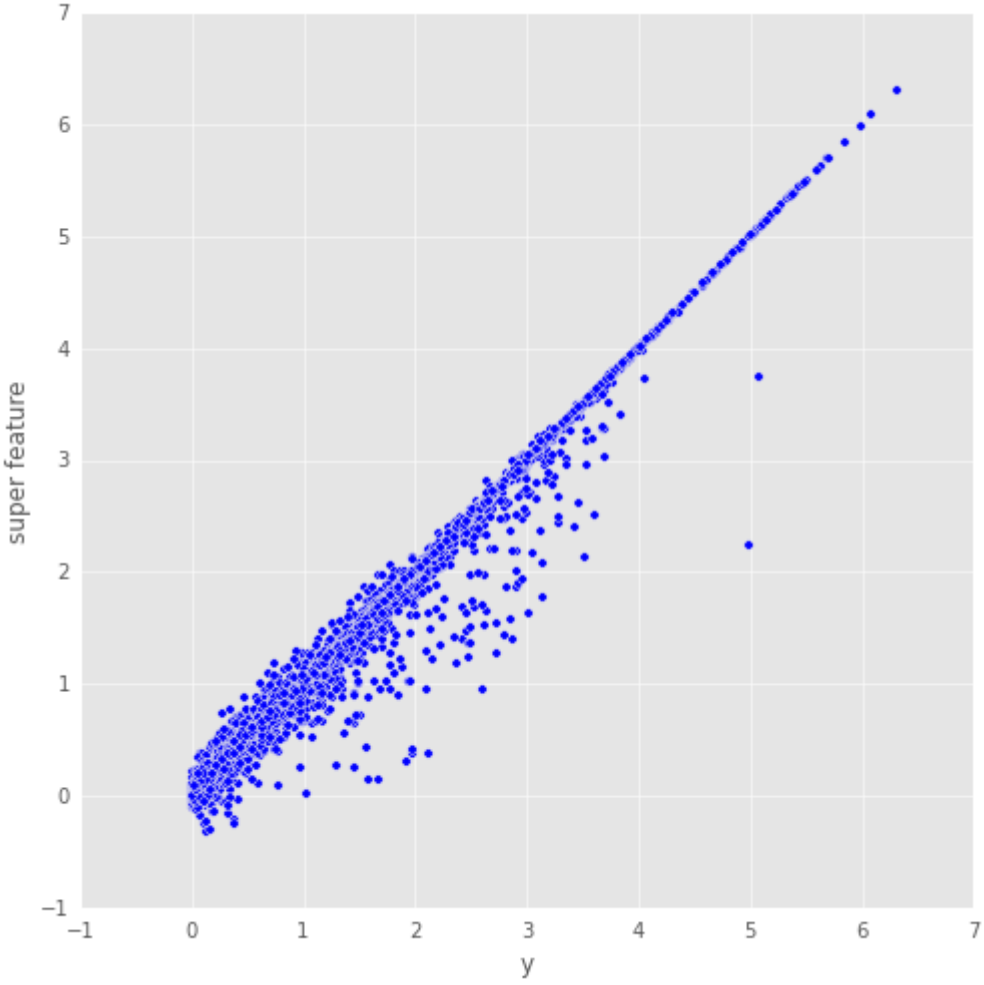
целевой признак и комбинация 2х признаков

Заметим, что эта комбинация строится как почти ответ…

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

Логарифмирование целевого признака

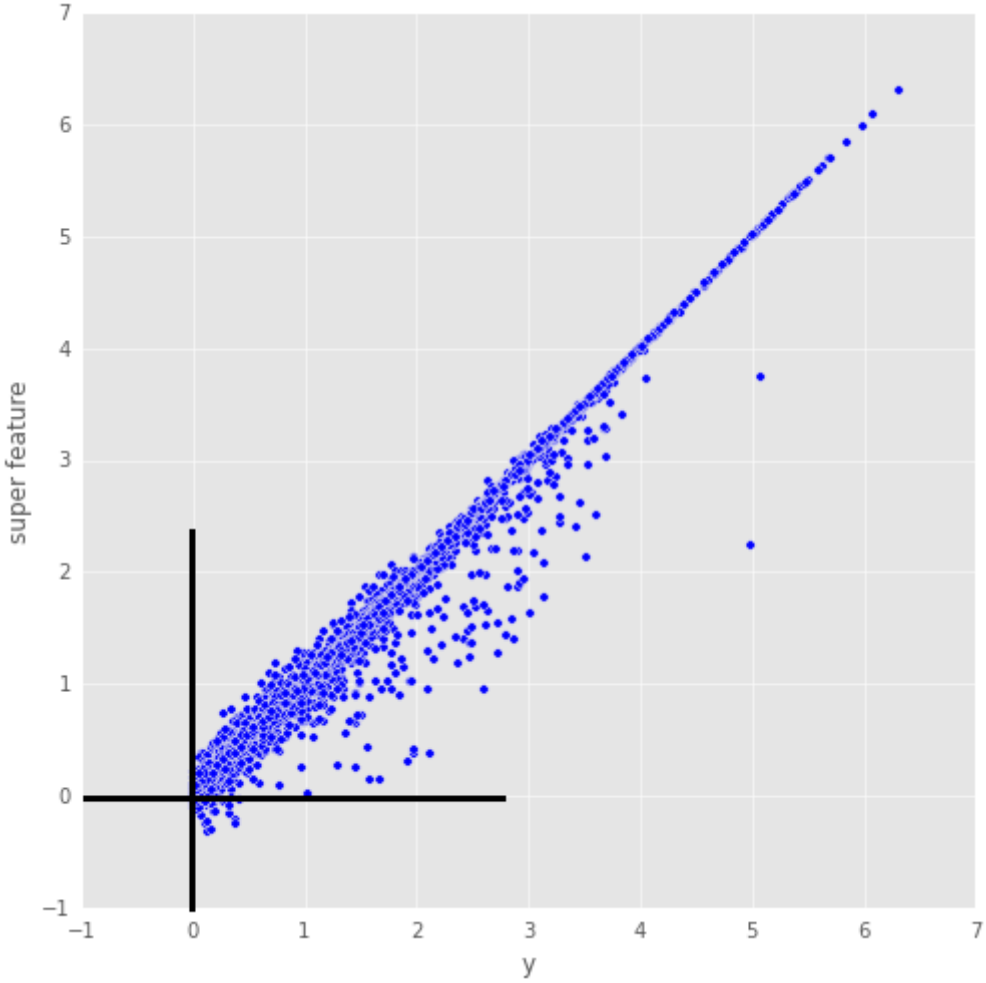
Что ещё сделать, чтобы картинка стала понятнее?



целевой признак и комбинация 2х признаков

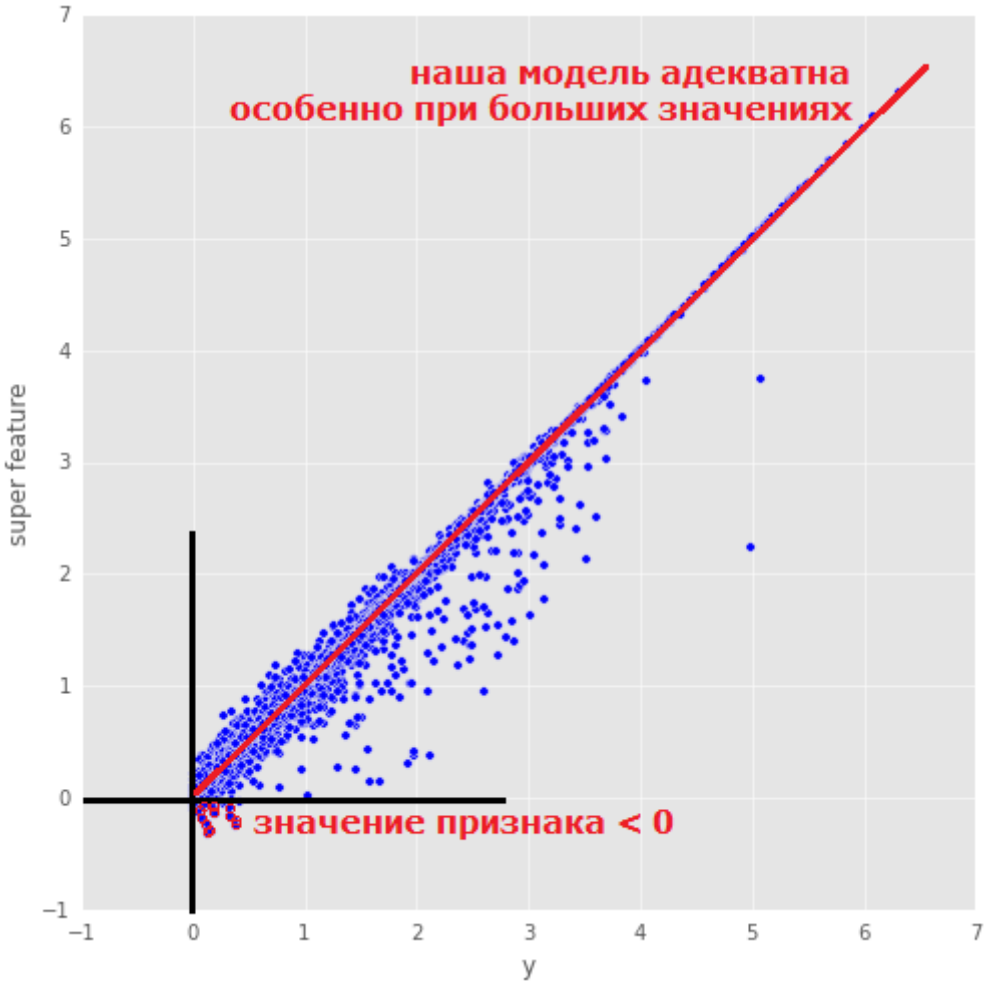
```
plt.scatter(np.log(y2), np.log(train2.mnk.values) + train2.tmp.values)
```

Логарифмирование целевого признака



Что видно на графике?

Логарифмирование целевого признака

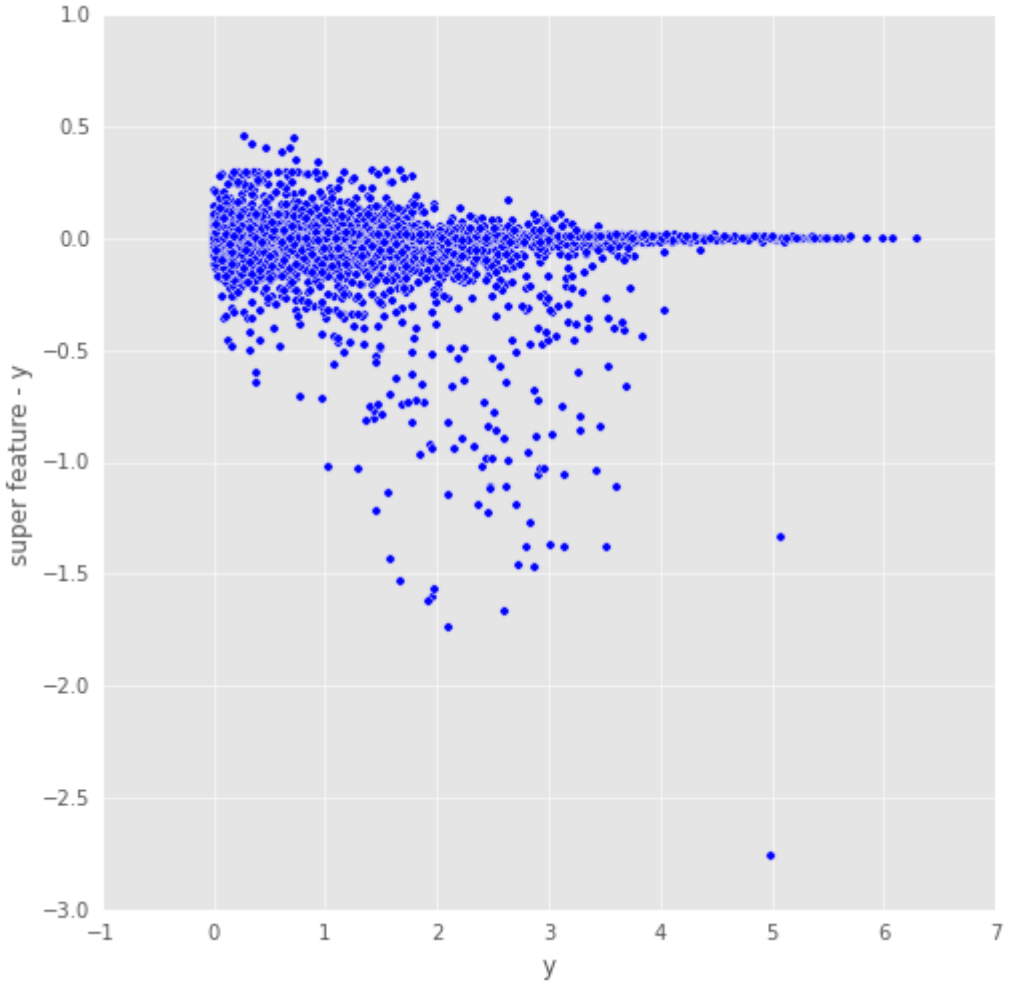


Правильный ответ всегда > 0

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

Вывод: `maximum(f, 0)`

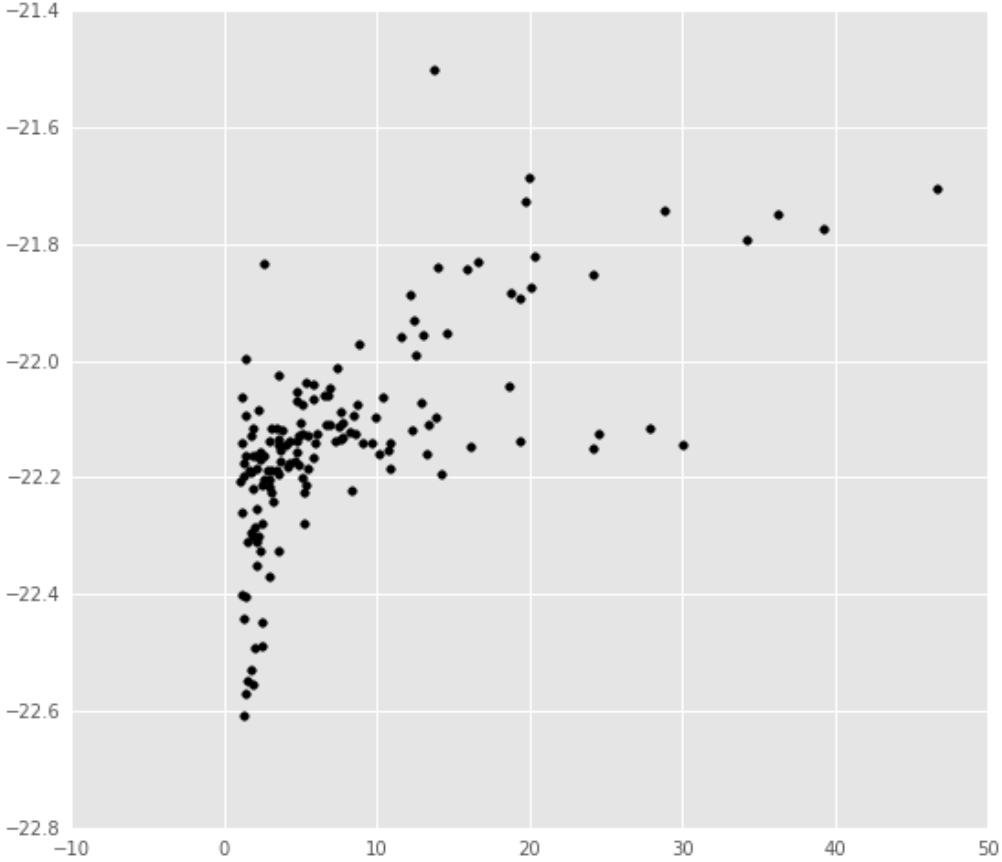
Разница признака и целевого признака



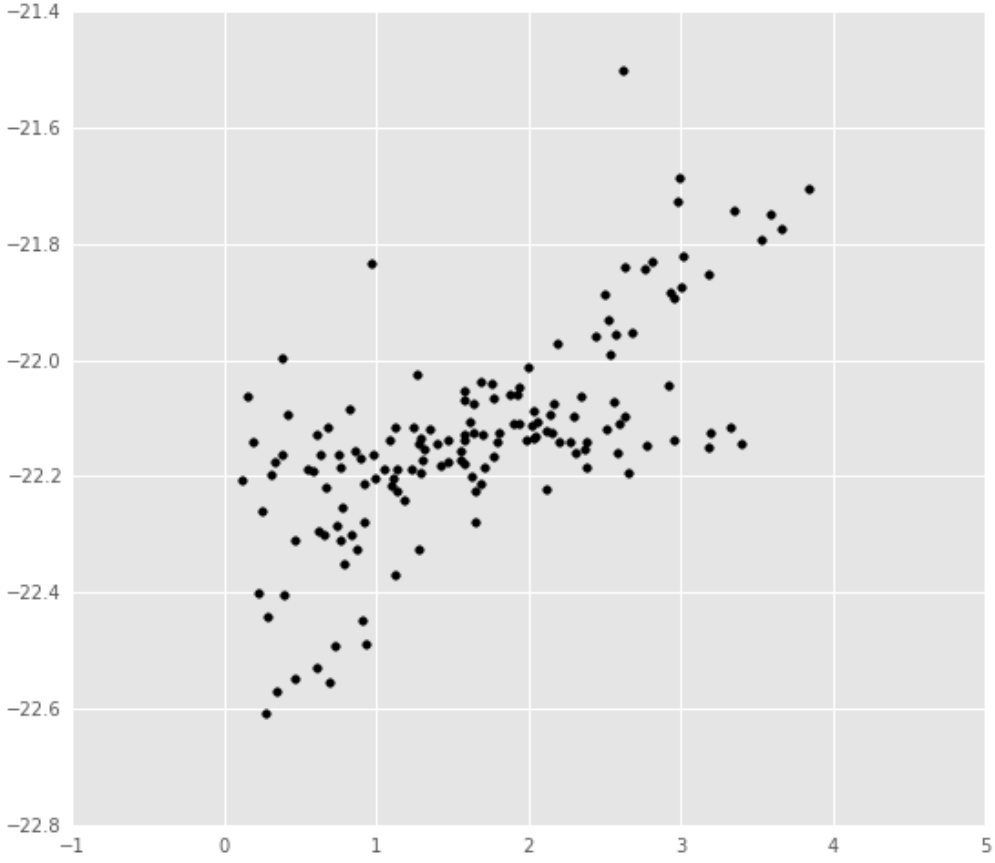
Если построили «почти ответ» – полезно посмотреть на ошибку

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

Необходимость логарифмирования можно не заметить на меленьких выборках

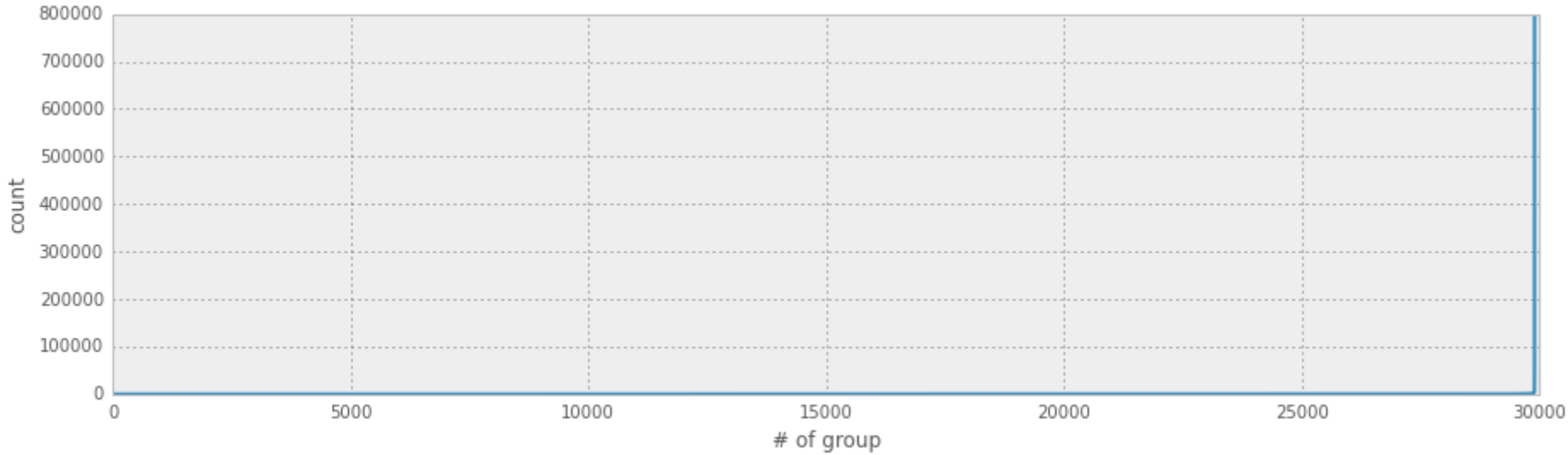


До логарифмирования

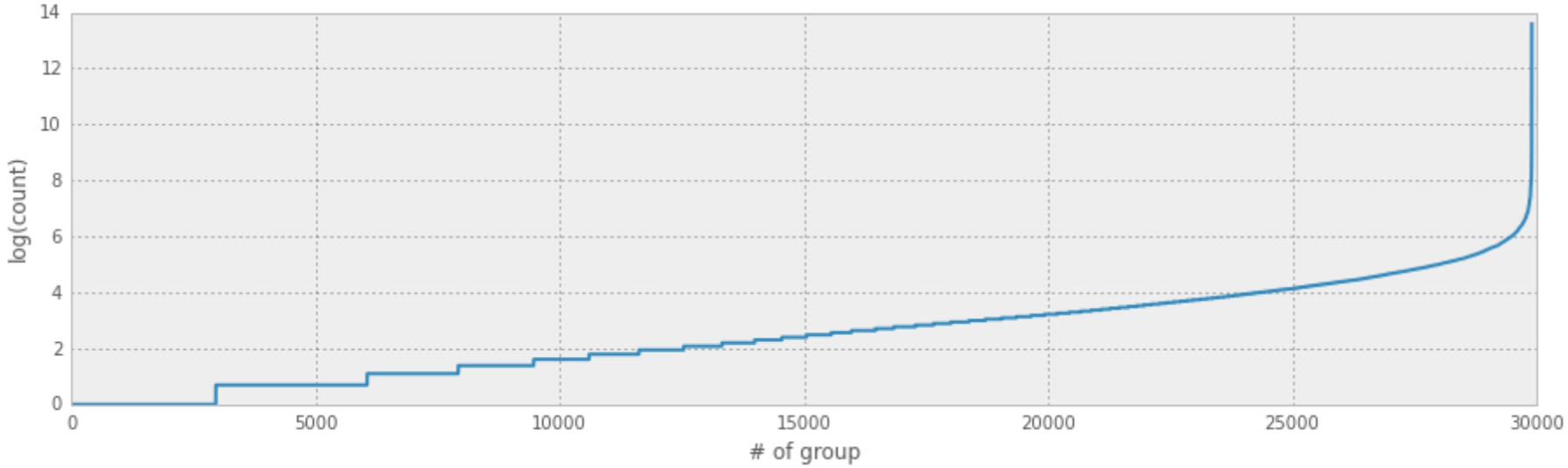


после

Зачем ещё нужно логарифмирование

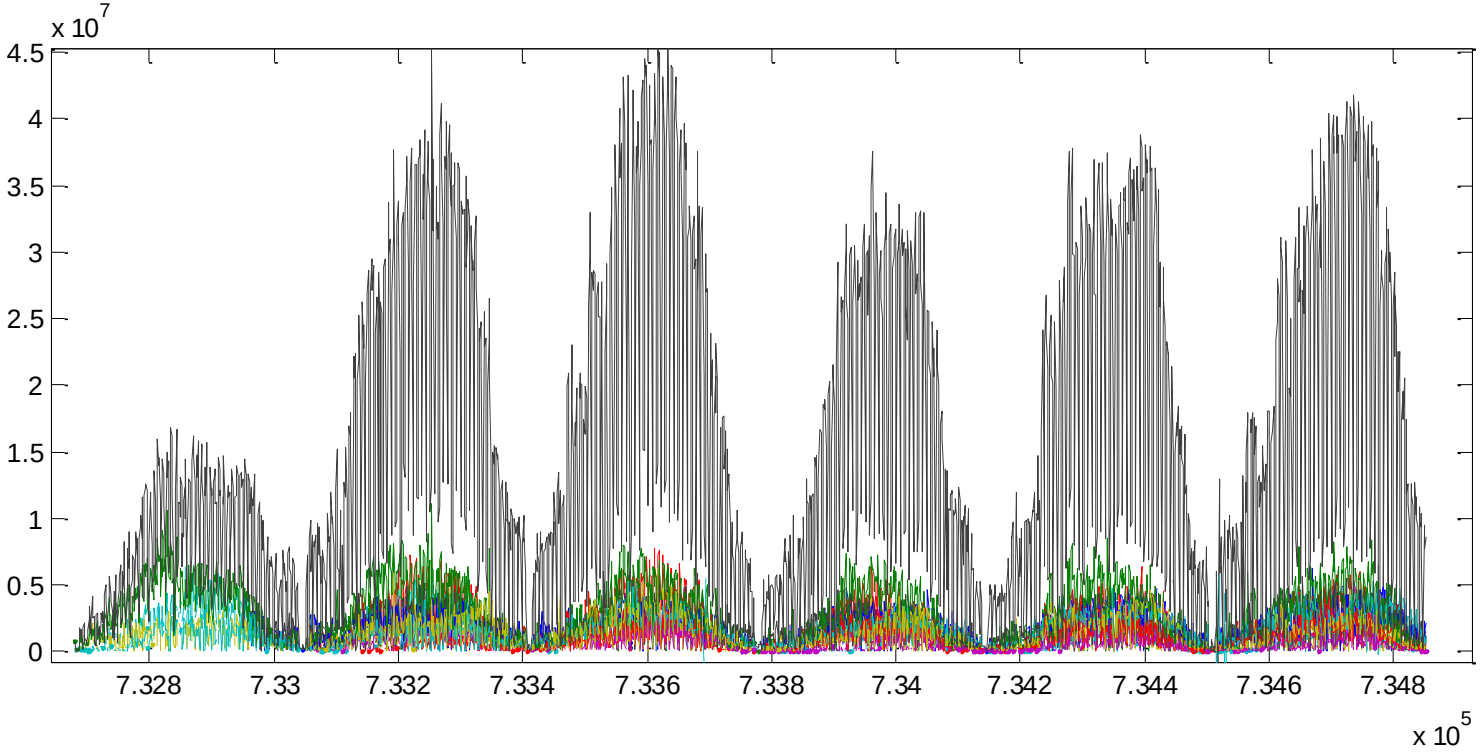


число представителей одной из ~30000 групп в выборке



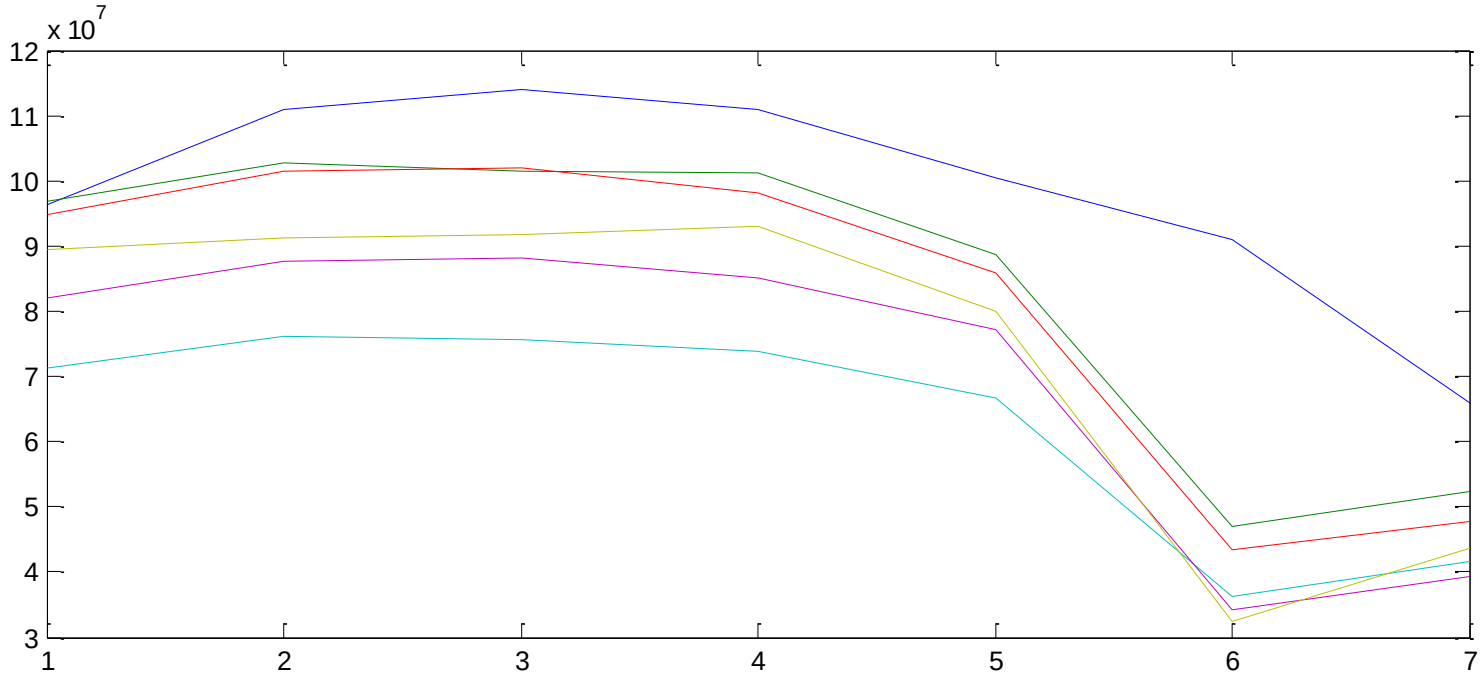
логарифм этого числа

Другая задача: прогнозирование временного ряда (продажи)



Есть отрицательные значения – выбросы вниз (!?).

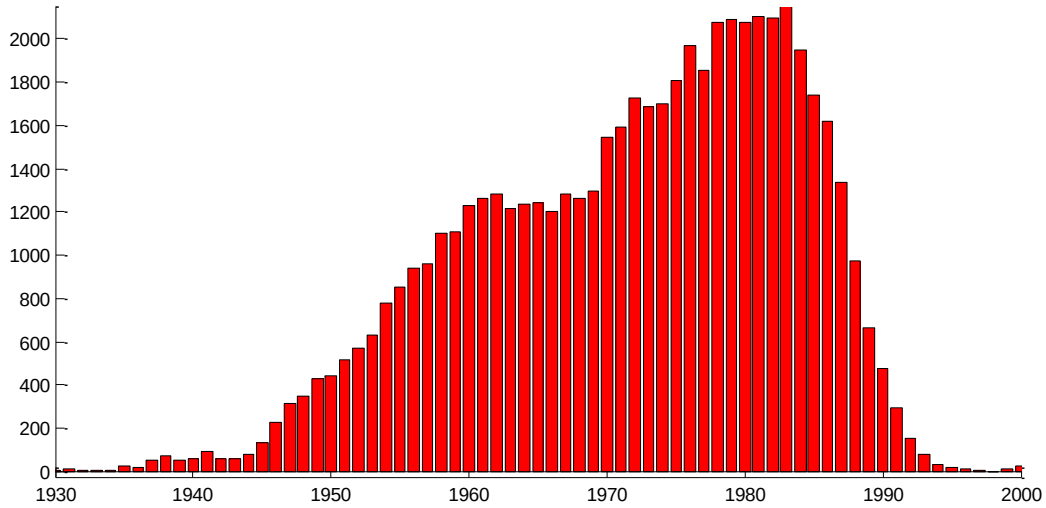
Если усреднить недели каждого года:



Первый год нетипичен!

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

ЗАДАЧА «М-магазин»

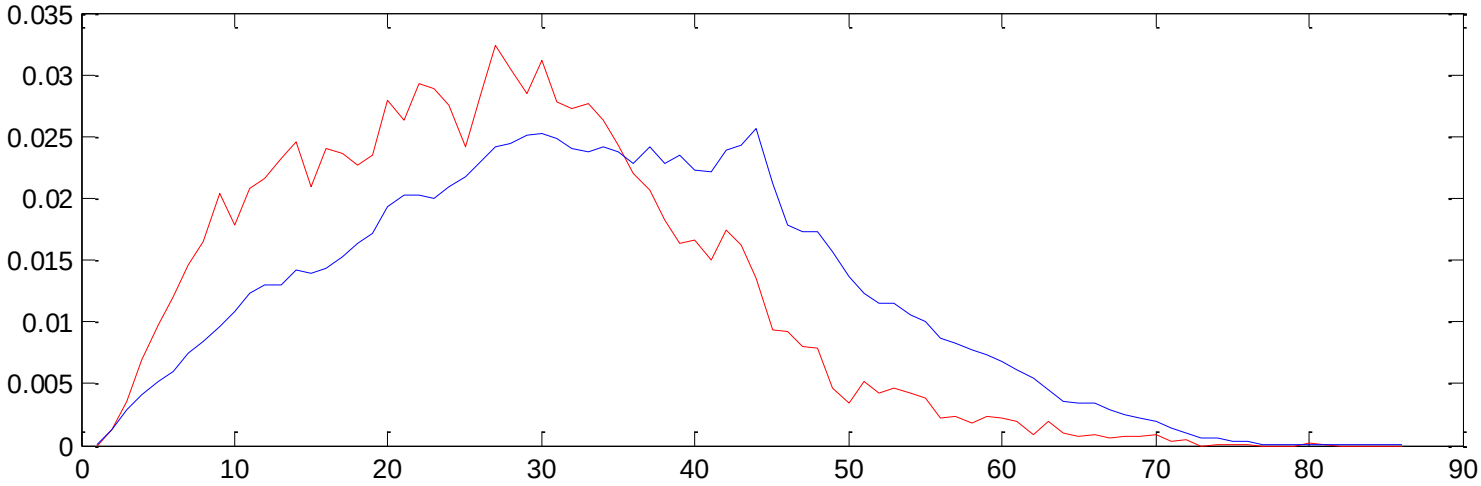


Распределение возраста покупателей

Так обычно выглядит распределение!

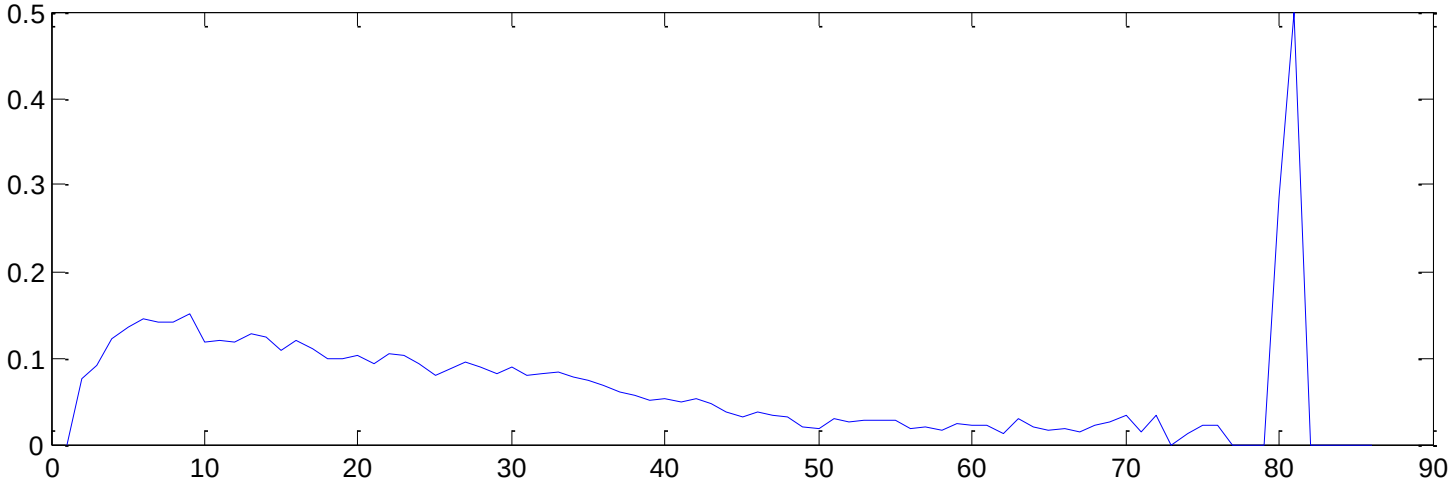
ЗАДАЧА «CREDIT»

AGE

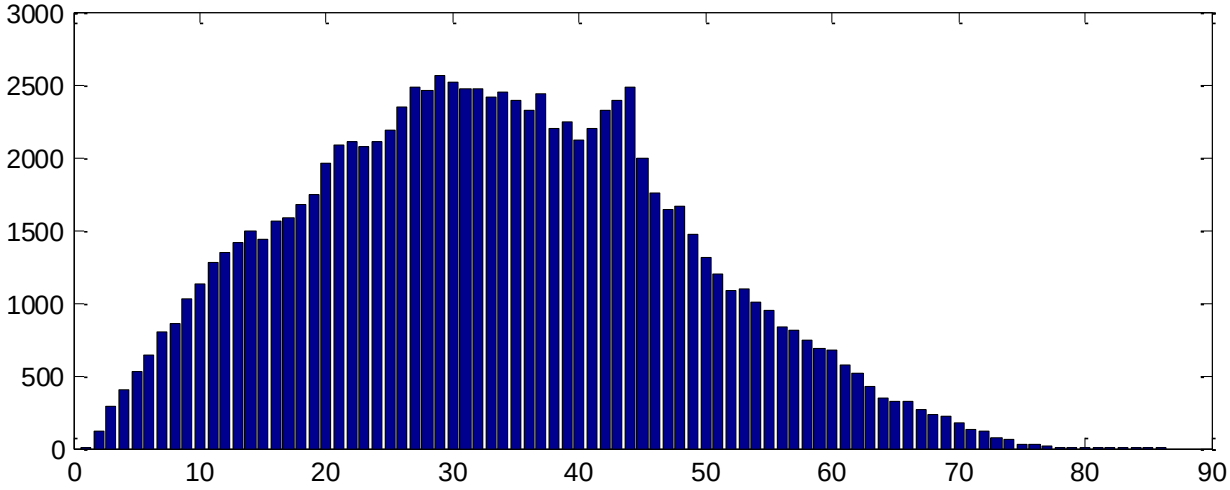


Плотности. Признак «возраст».

Случай из жизни: цена + страховка.

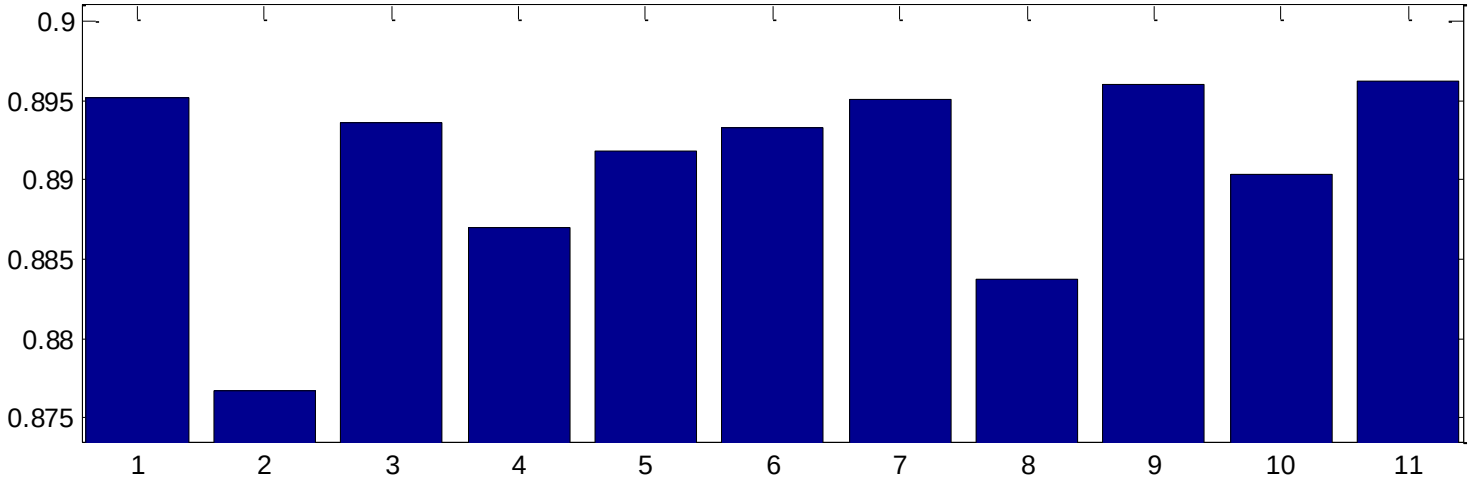


Отношение плотностей – есть явный выброс!



Распределение по возрасту

Качества признаков



Выбирается метод (ex: RF).

Первый столбец – качество на всех признаках, а потом – при удалении отдельных признаков.

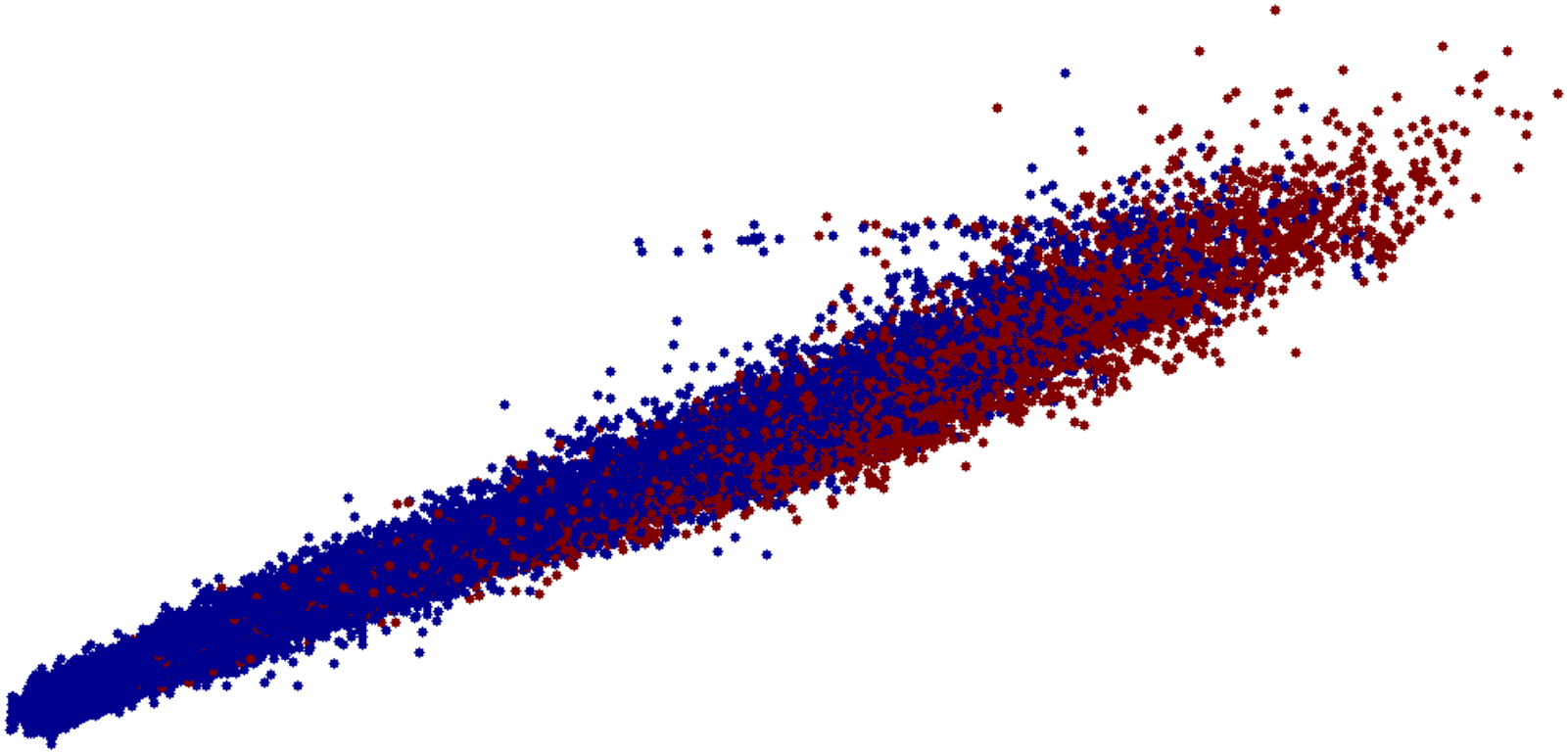
Есть более интересные методы

1. Importance

2. Boruta

Ещё о визуализации «алгоритм-алгоритм»

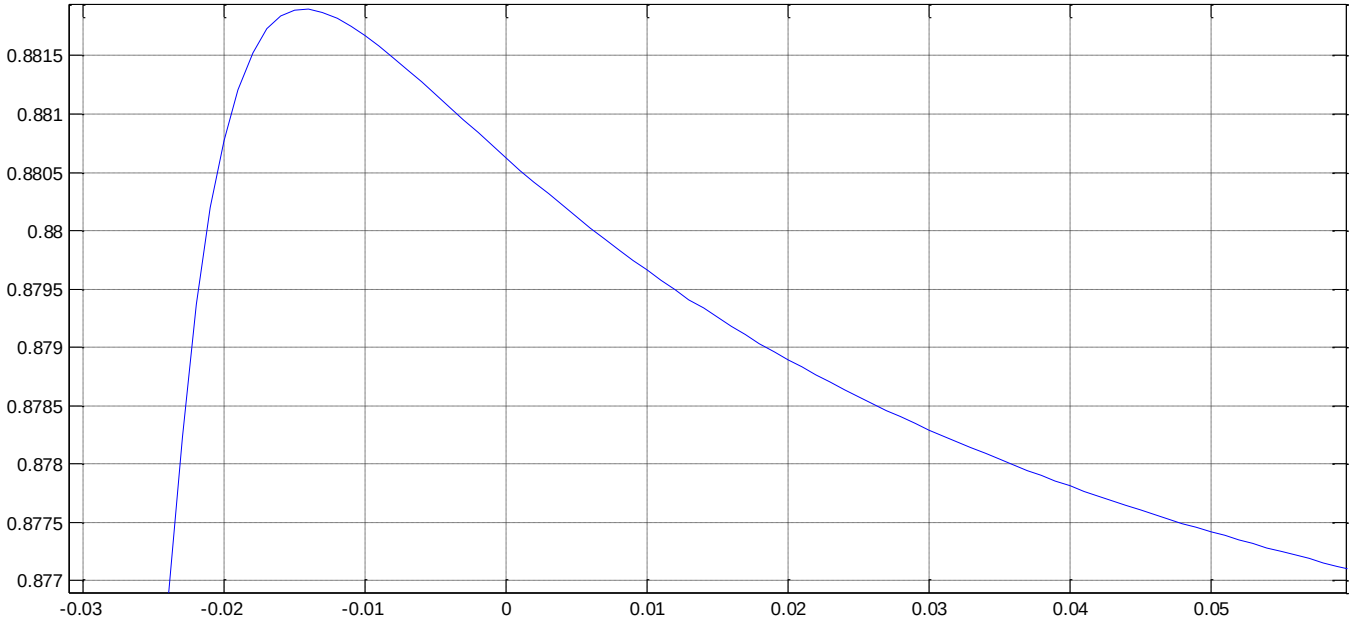
Задача скоринга



Байес и (RF+GBM)

Меньше похоже на отрезок и сигмоиду

Анализ коэффициента в выпуклой комбинации

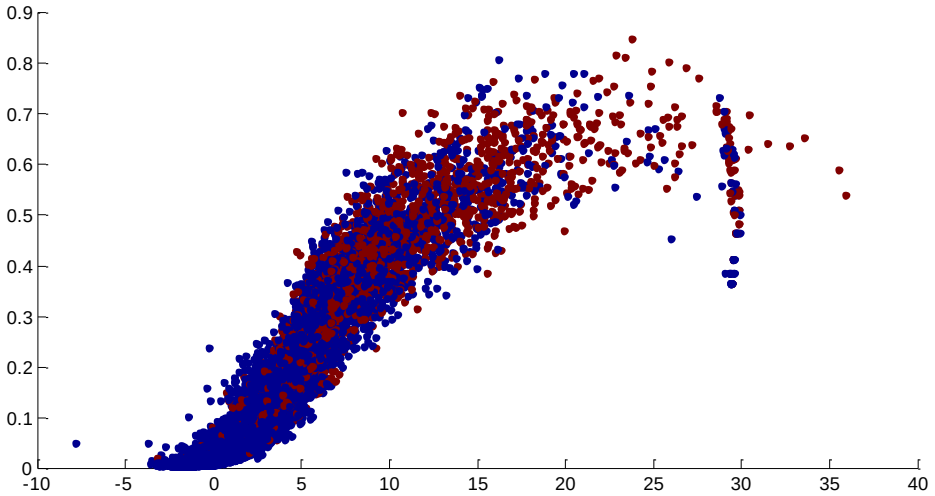


Коэффициент в линейной комбинации.

Лучше вычитать!

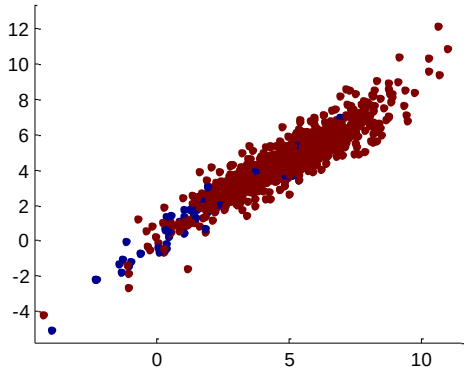
Ещё о визуализации «алгоритм-алгоритм»

Задача скоринга

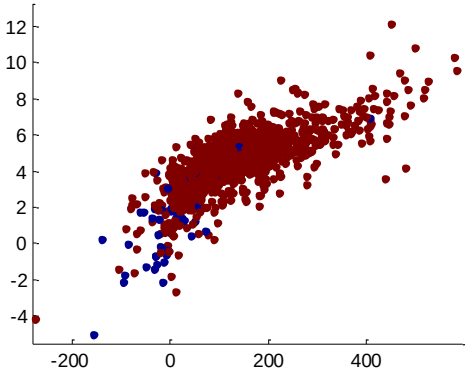


Мой – горизонталь и RF – вертикаль

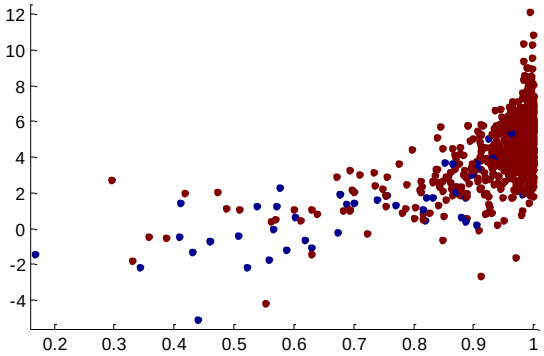
В задаче AMAZON



Разные LIBLINEAR



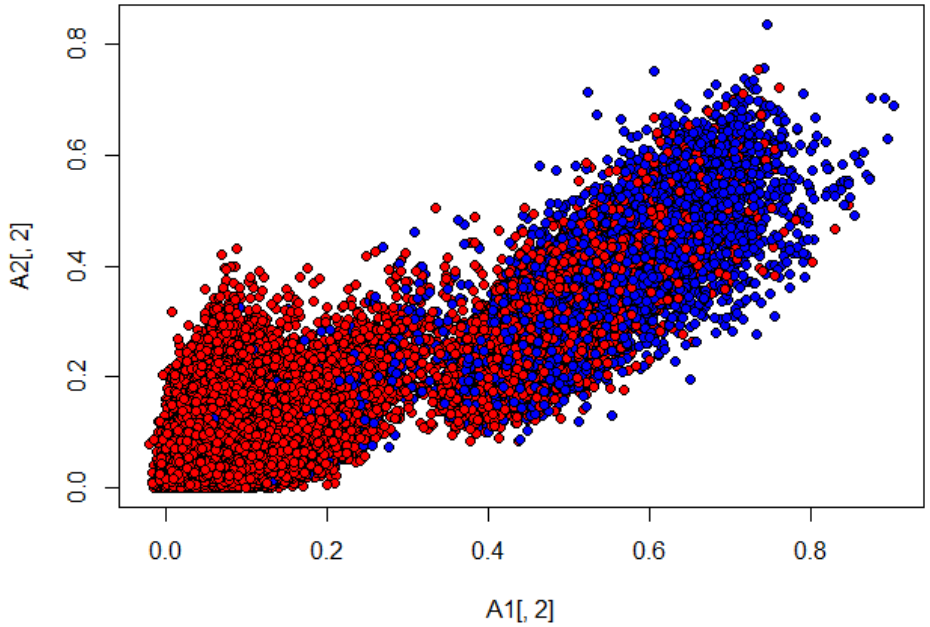
LIBLINEAR и PERCEPTRON

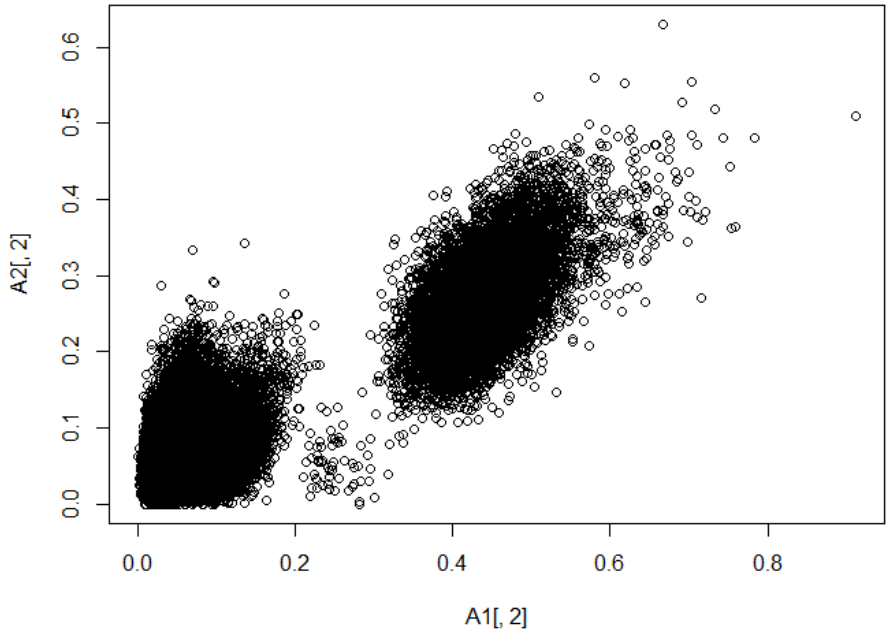


kNN и LIBLINEAR

Ещё о визуализации «алгоритм-алгоритм»

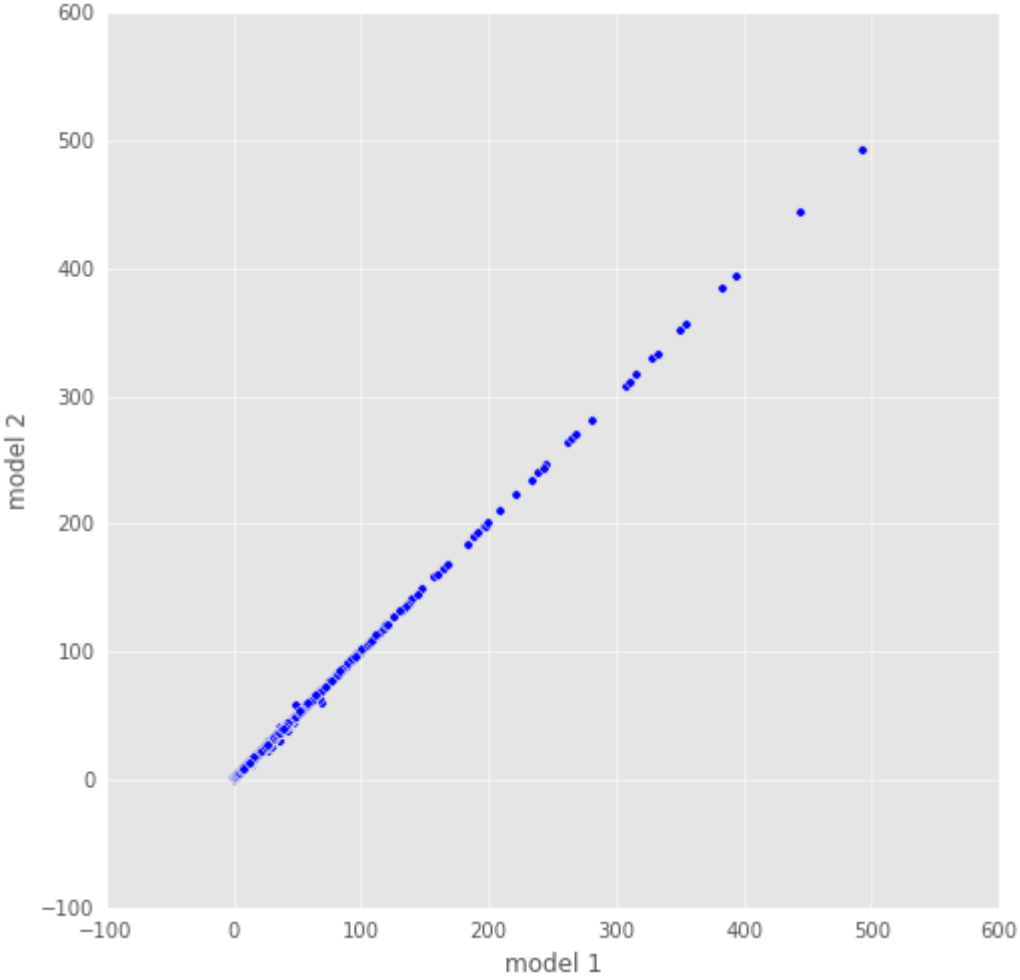
История обнаружения одной ошибки: бенчмарк – финальный алгоритм



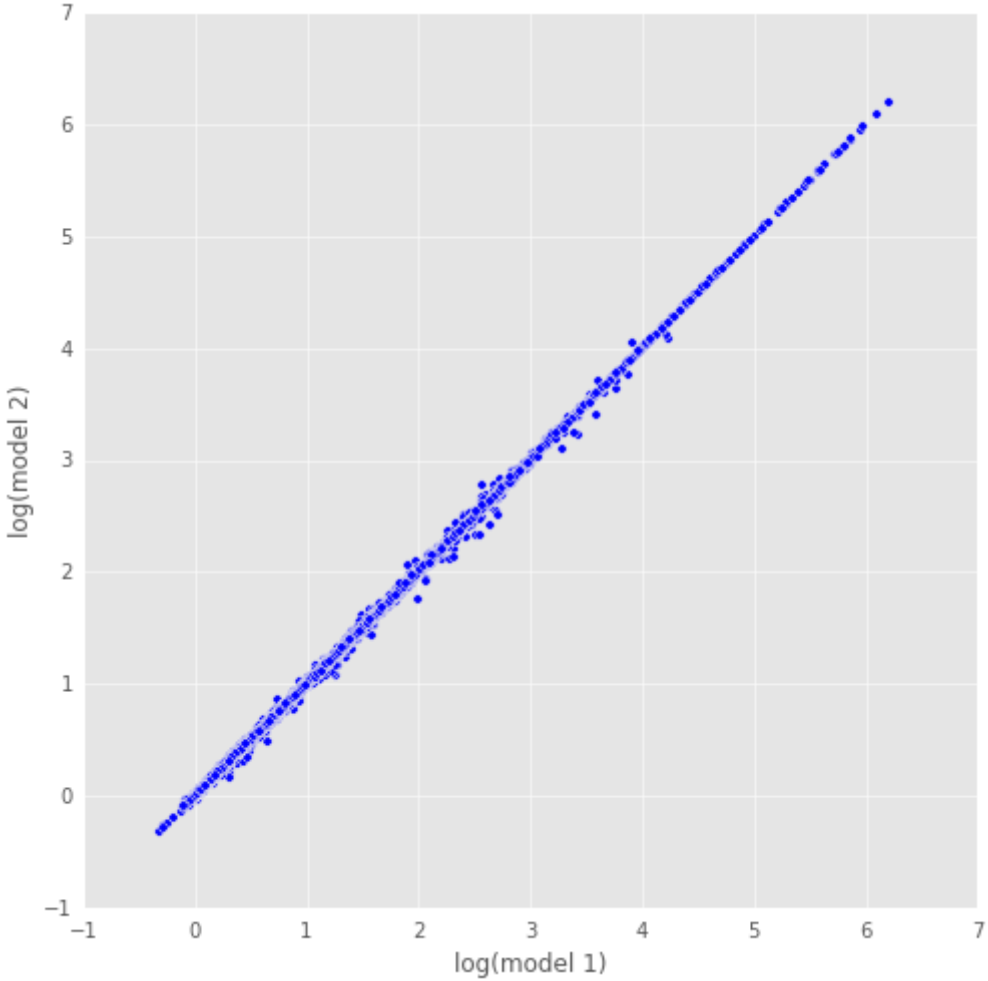


Смотрите пары решений на локальном и окончательном контроле

Ещё о визуализации «алгоритм-алгоритм»

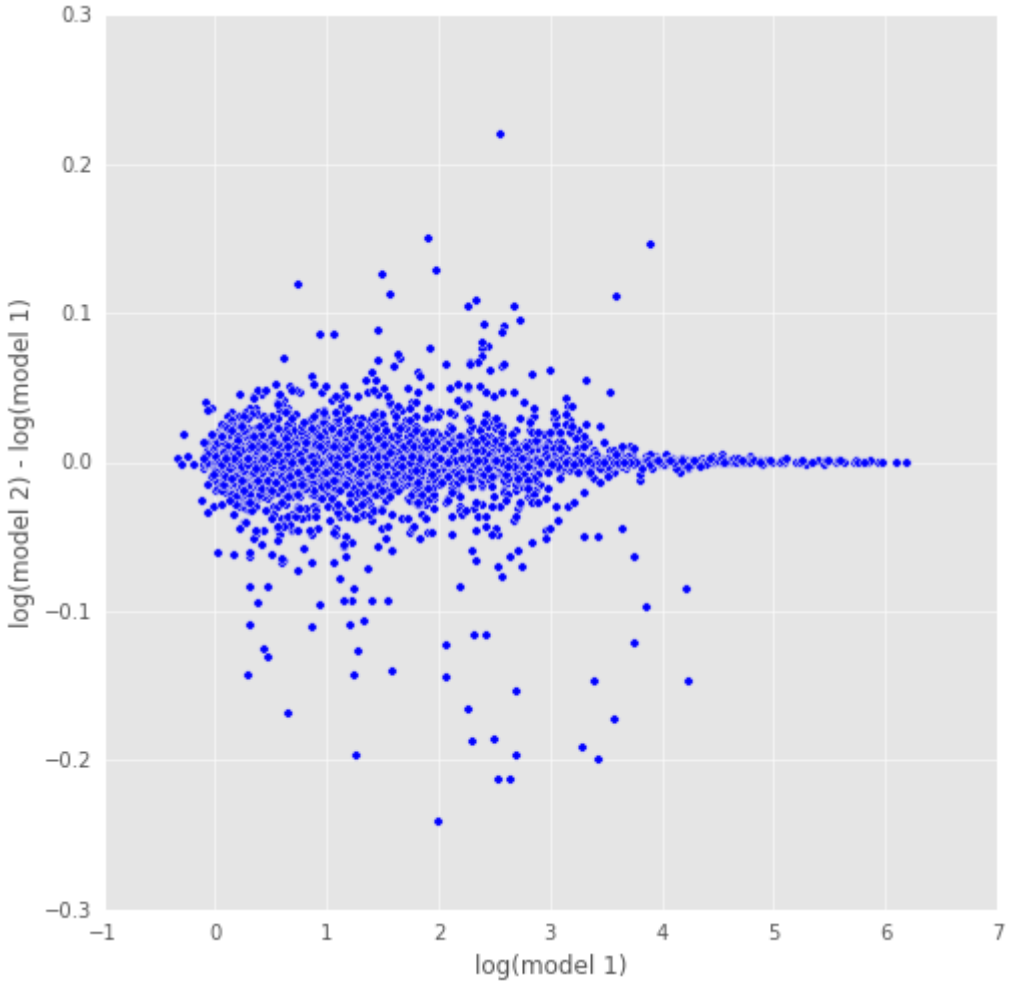


Две модели



Опять логарифмирование шкал!

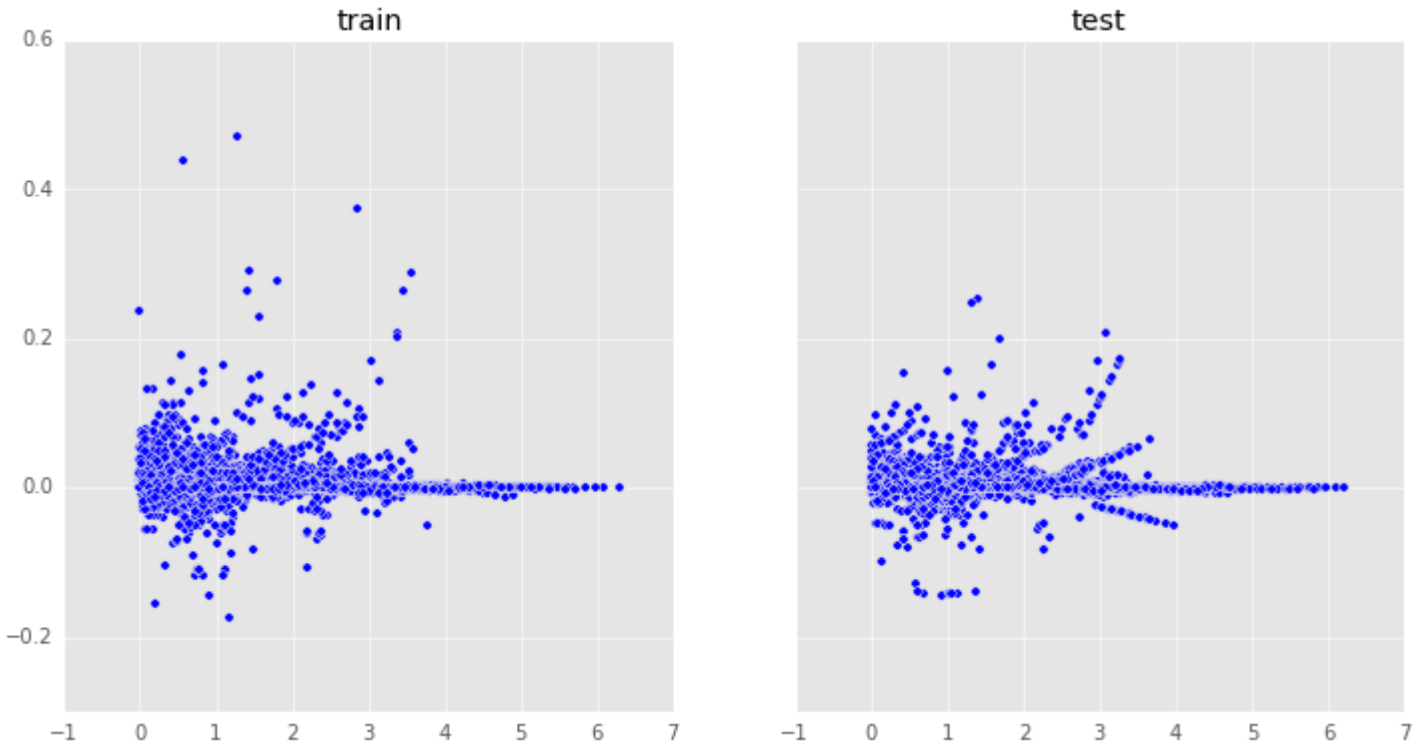
Ещё о визуализации «алгоритм-алгоритм»



Опять смотрим разницу ответов

Наблюдение: при больших значениях модели работают идентично!

Ещё о визуализации «алгоритм-алгоритм»



На контроле подозрительные линии…

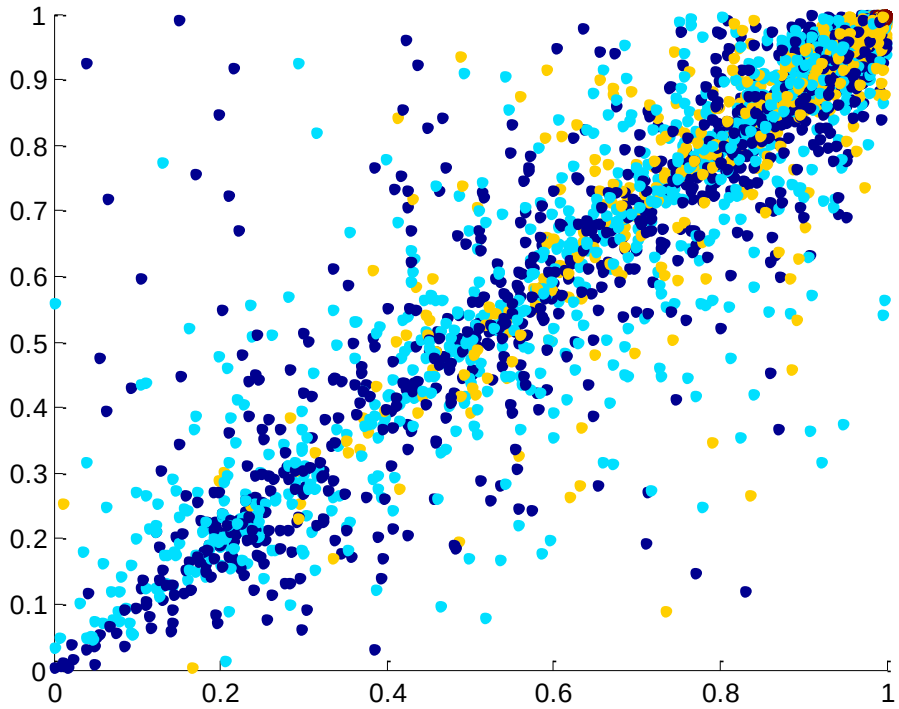
Что это может значить?

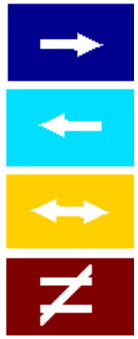
Что делать?

Задача «Причина-следствие»

Метод: «ручная деформация пространств»

```
% метод, основанный на полиномиальном приближении
[f fn] = cause_f_polyfit(Xs);
scatter(f(:,1), f(:,2), 20, Ys(:,2), 'filled')
```

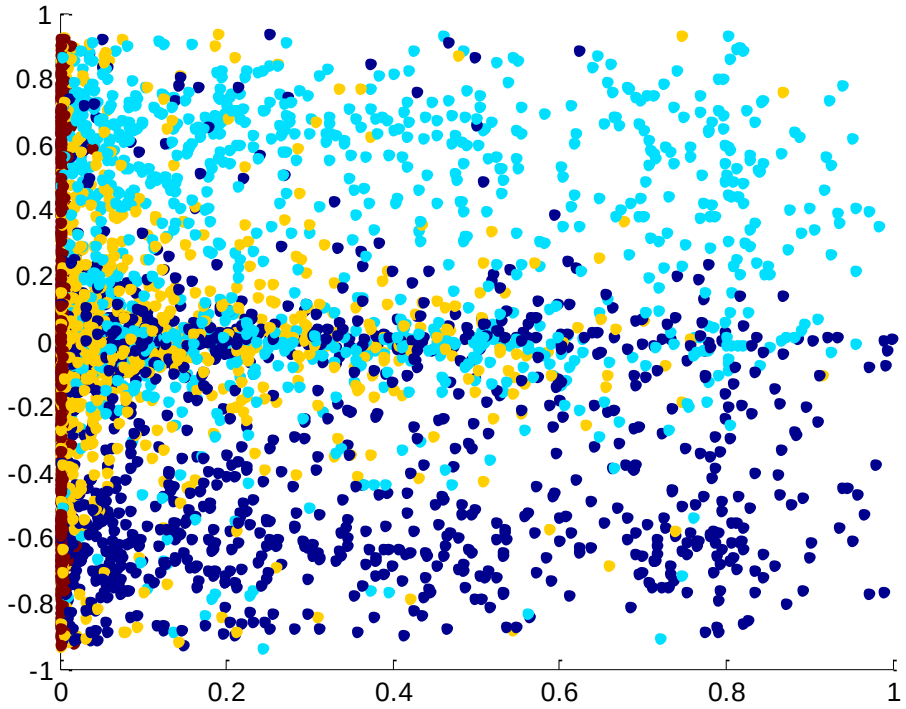




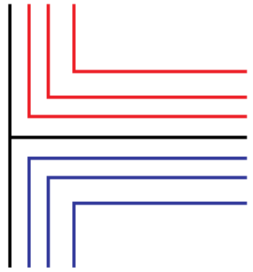
Кстати: хорошая задача – пример «новой науки»

Алгебраические выражения над признаками

```
scatter(1-0.5*(f(:,1)+f(:,2)),fn21(:,1)-fn21(:,2), 20, Ys(:,2), 'filled')
```

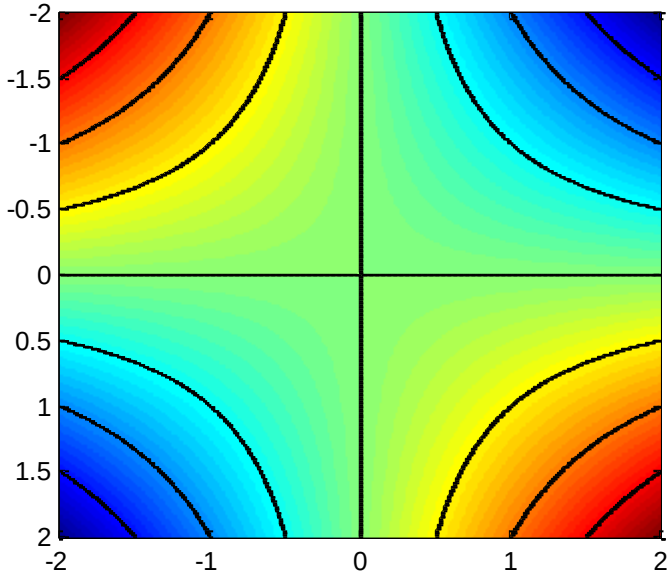


А теперь надо «уголками откусывать классы»:

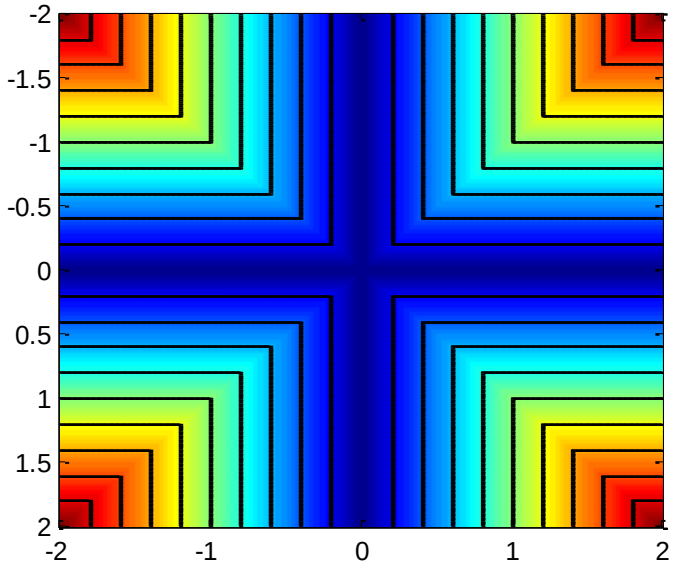


Какие функции «откусывают уголками»

$$z = y \cdot x$$

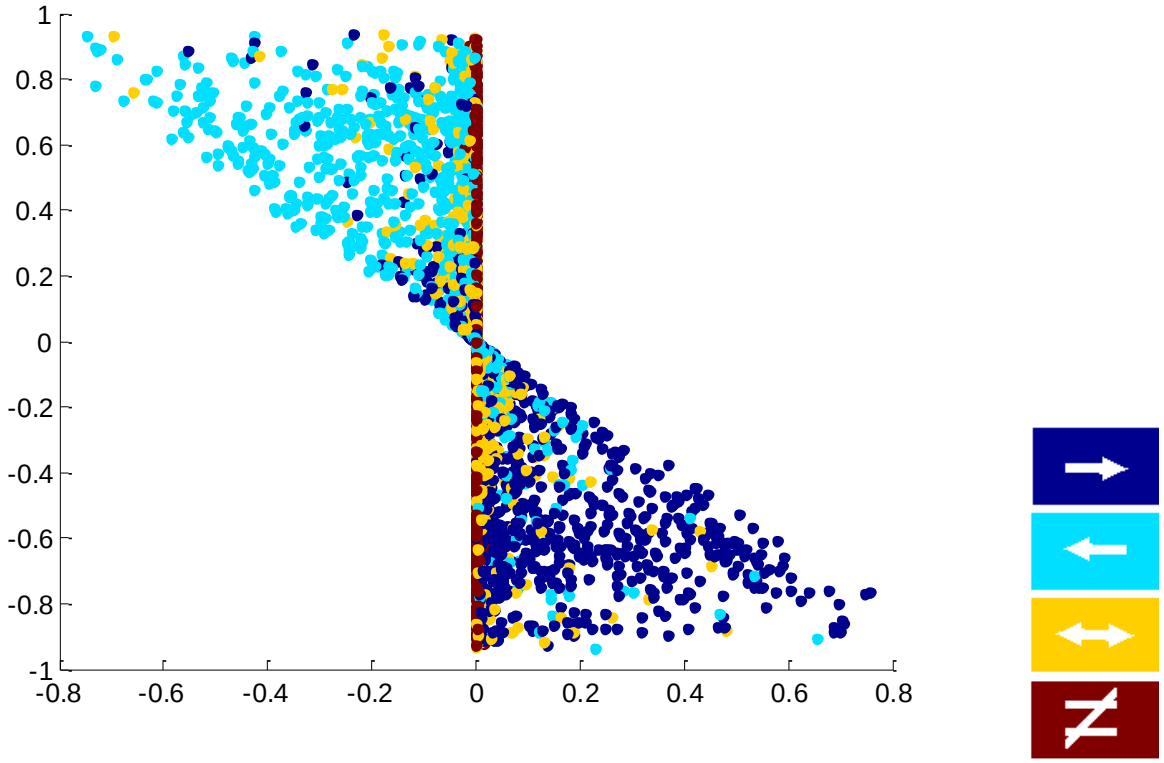


$$z = \min(|y|,|x|)$$



Алгебраические выражения над признаками

```
a = -(1-0.5*(f(:,1)+f(:,2))).*(fn21(:,1)-fn21(:,2))
scatter(a,fn21(:,1)-fn21(:,2), 20, Ys(:,2), 'filled')
```

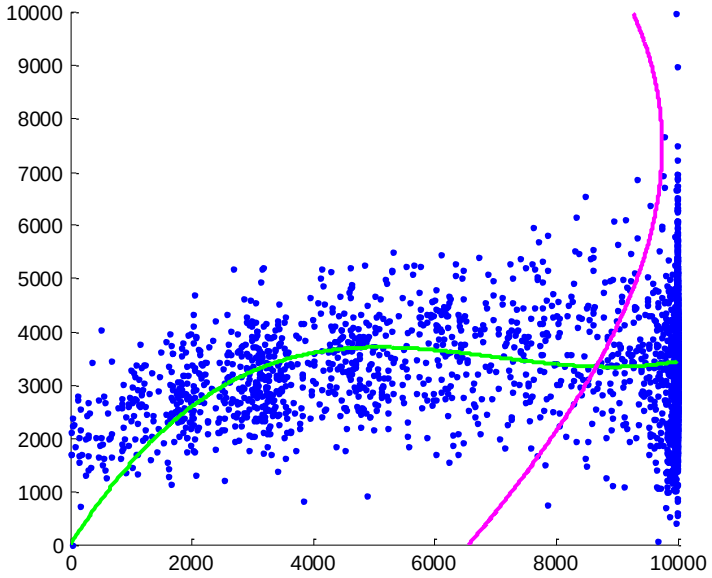


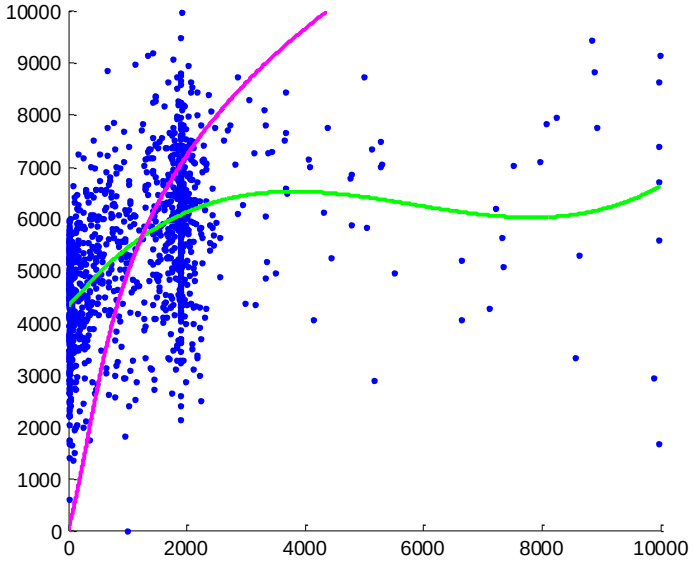
И здесь мы видим разделяемость синих и голубых!

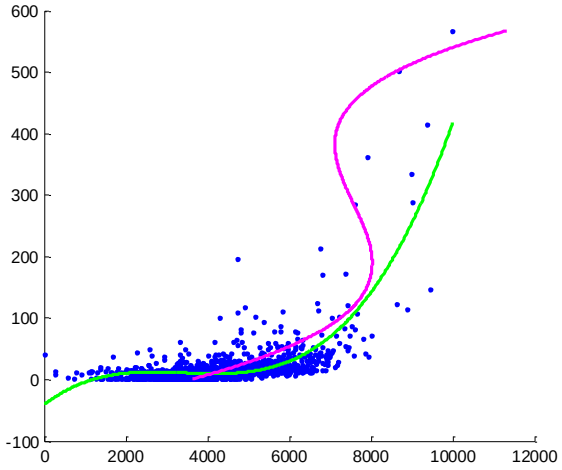
Получается алгоритм неплохого качества.

Ещё один приём: посмотреть как метод «работает»

Полиномиальная регрессии (deg=3) сразу от 2х переменных…

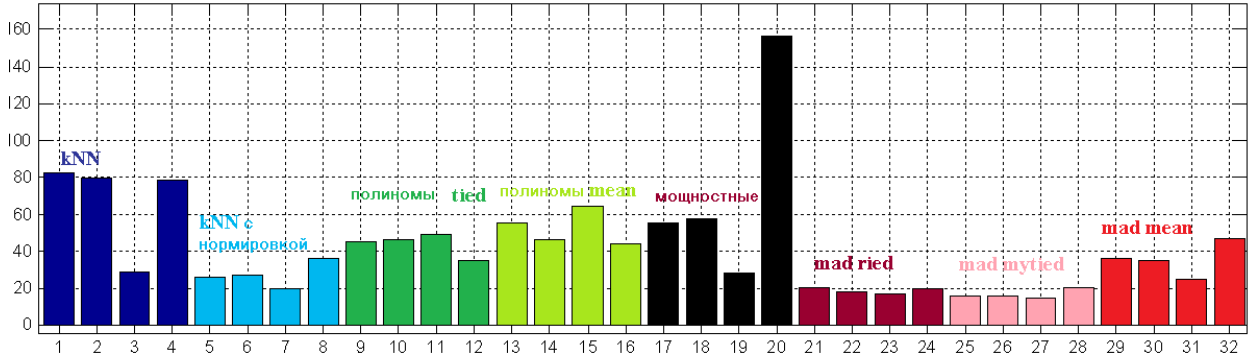




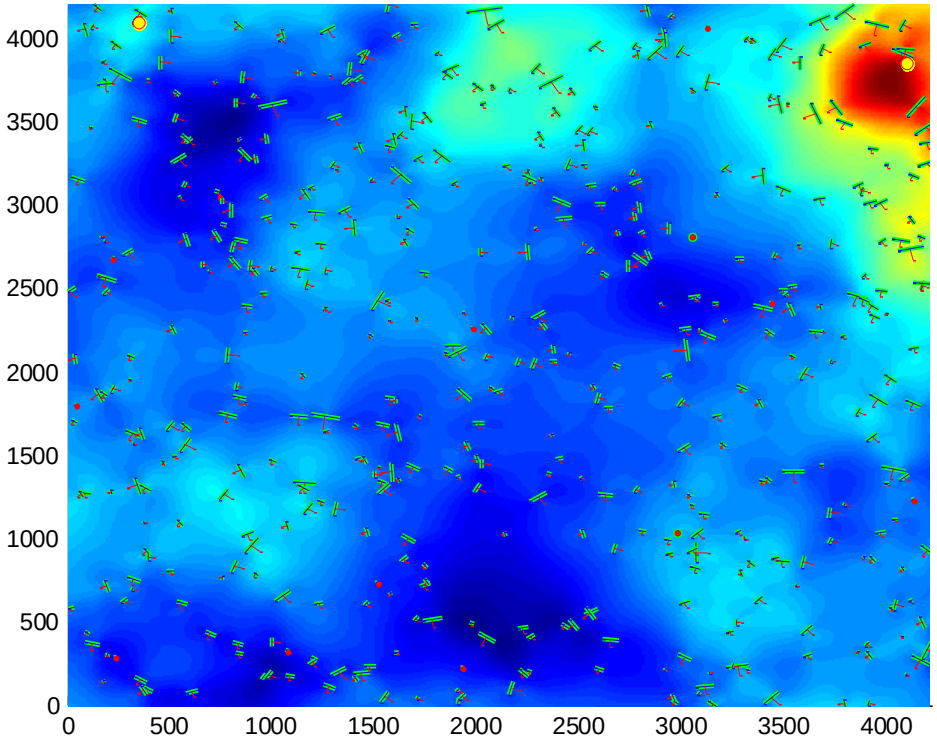


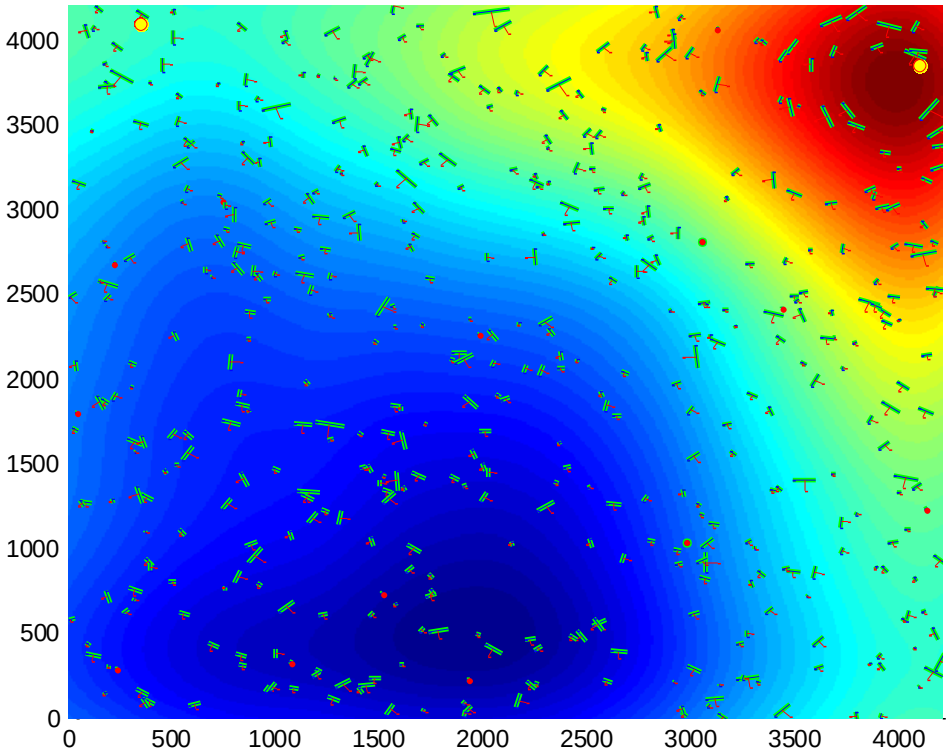
Ответы алгоритмов – как признаки

Построено несколько методов –
их ответы как признаки,
потом с помощью RF «качество алгоритмов».



Задача про чёрные дыры



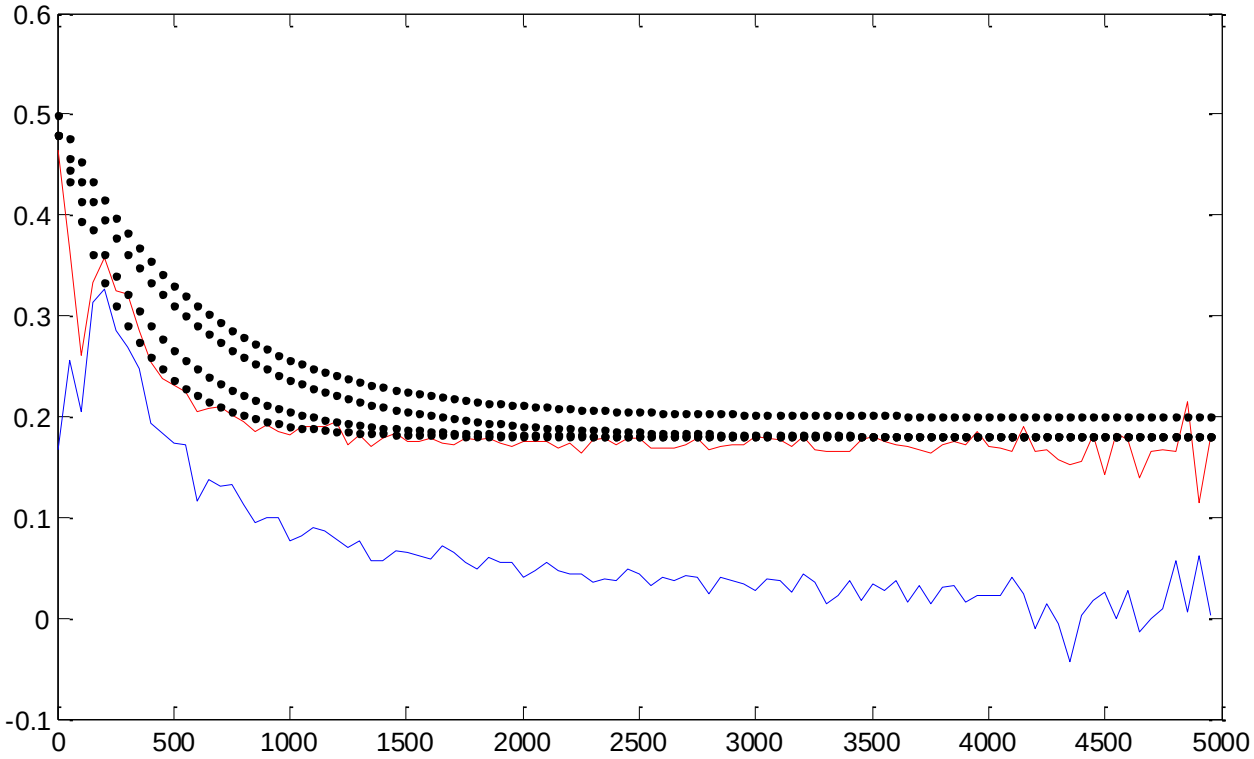


Какая связь между рисунками?

Ответ:

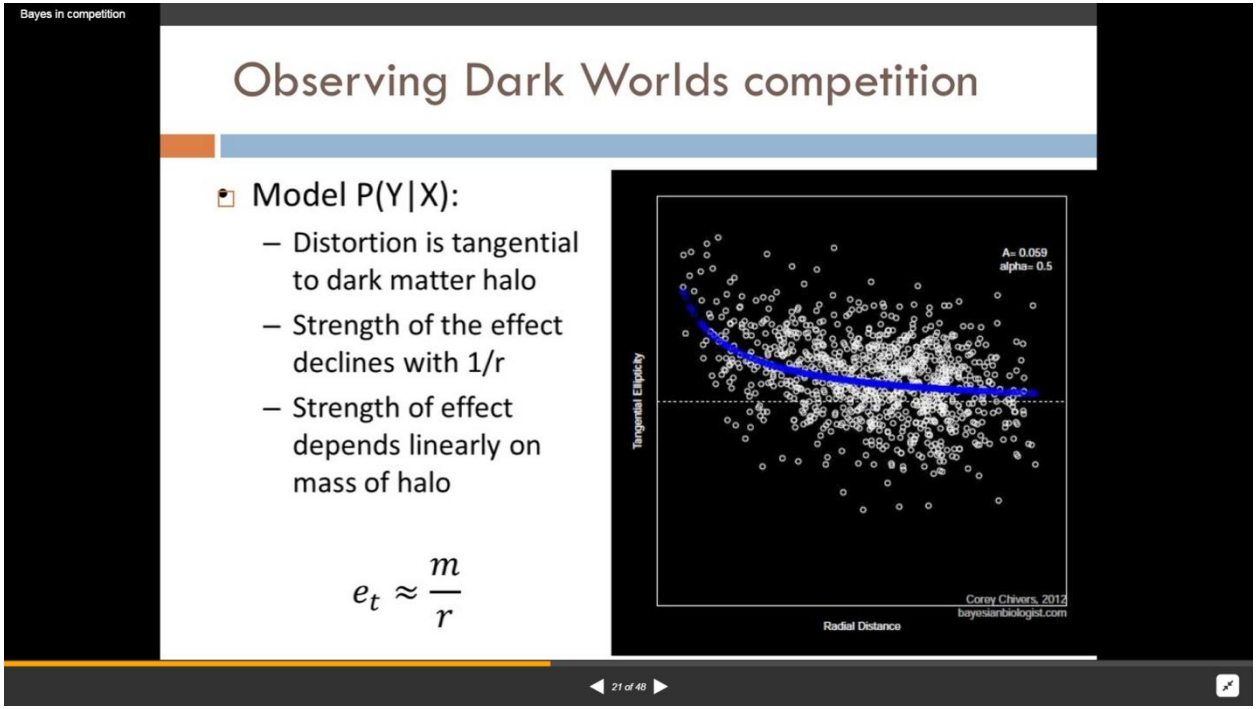
«Плотность» и её сглаженный аналог.

Средний профиль плотности(красный):



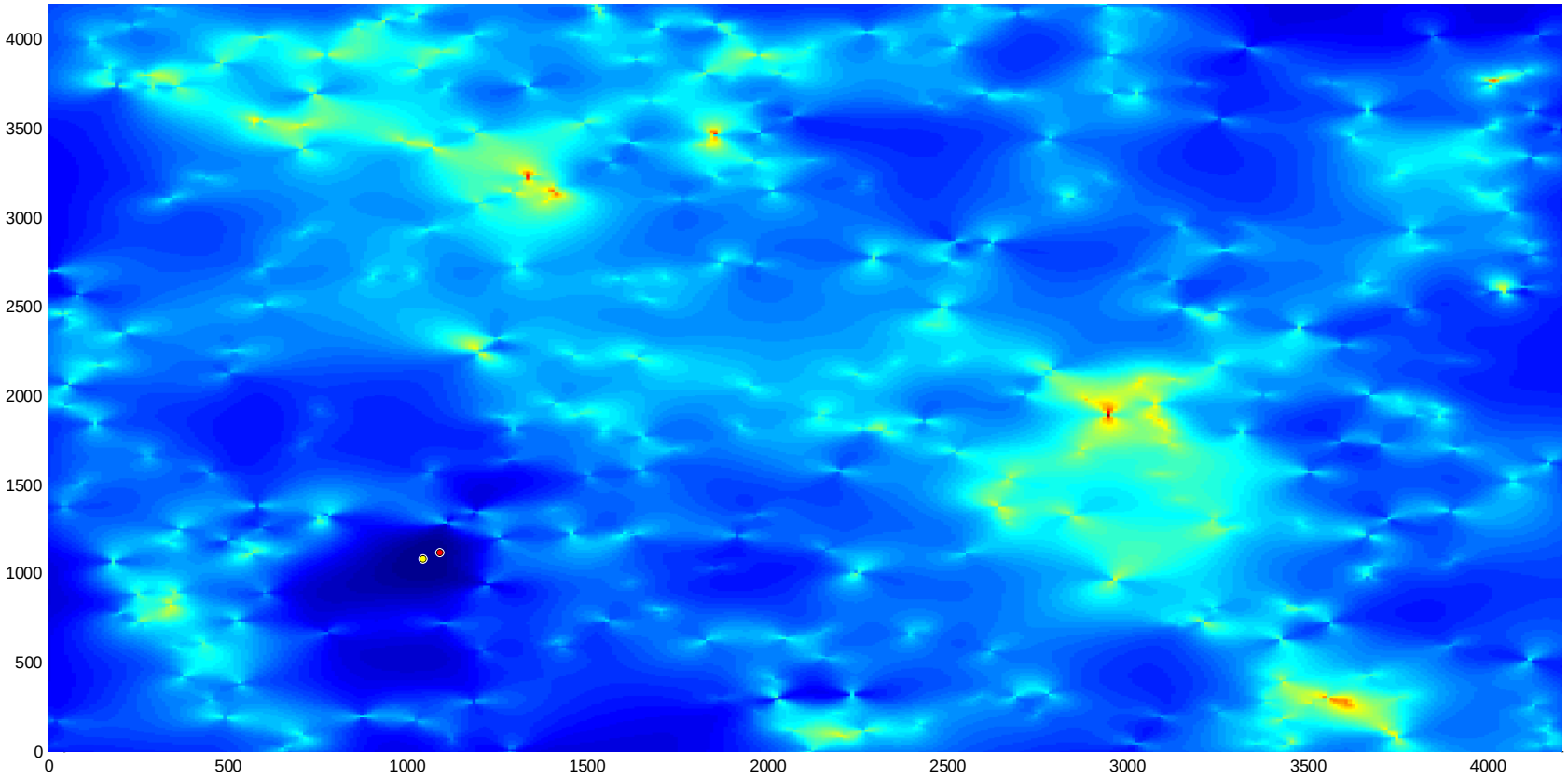
и методы его приближения

Owen Zhang



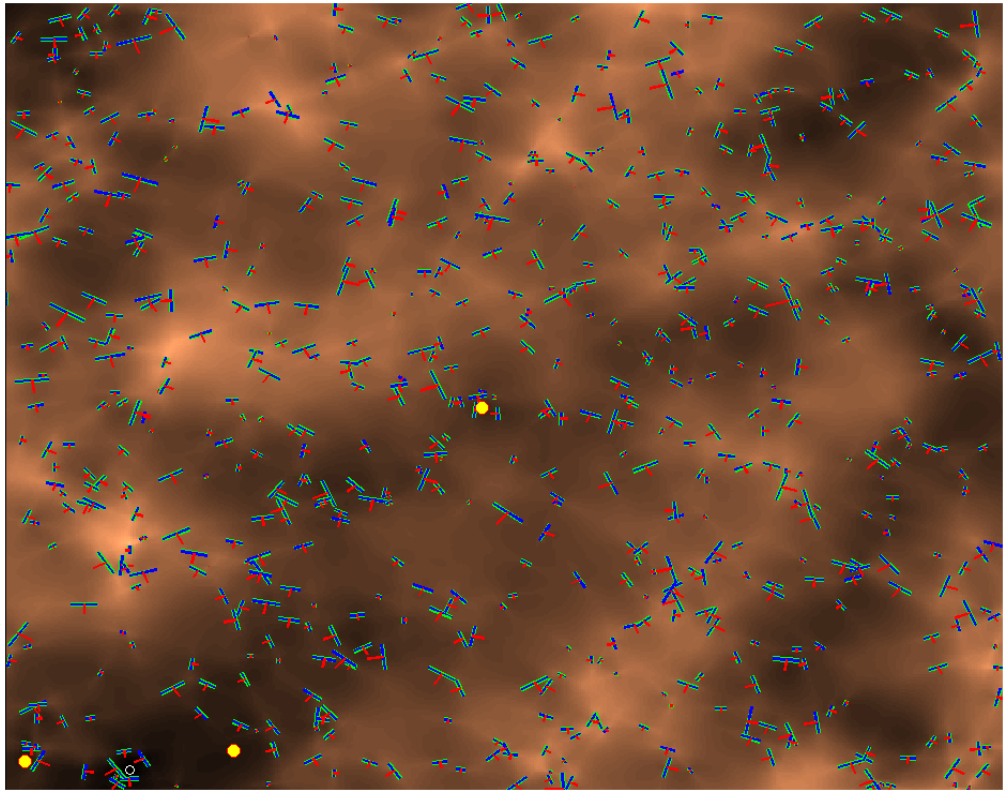
Также использовал визуализацию для создания модели

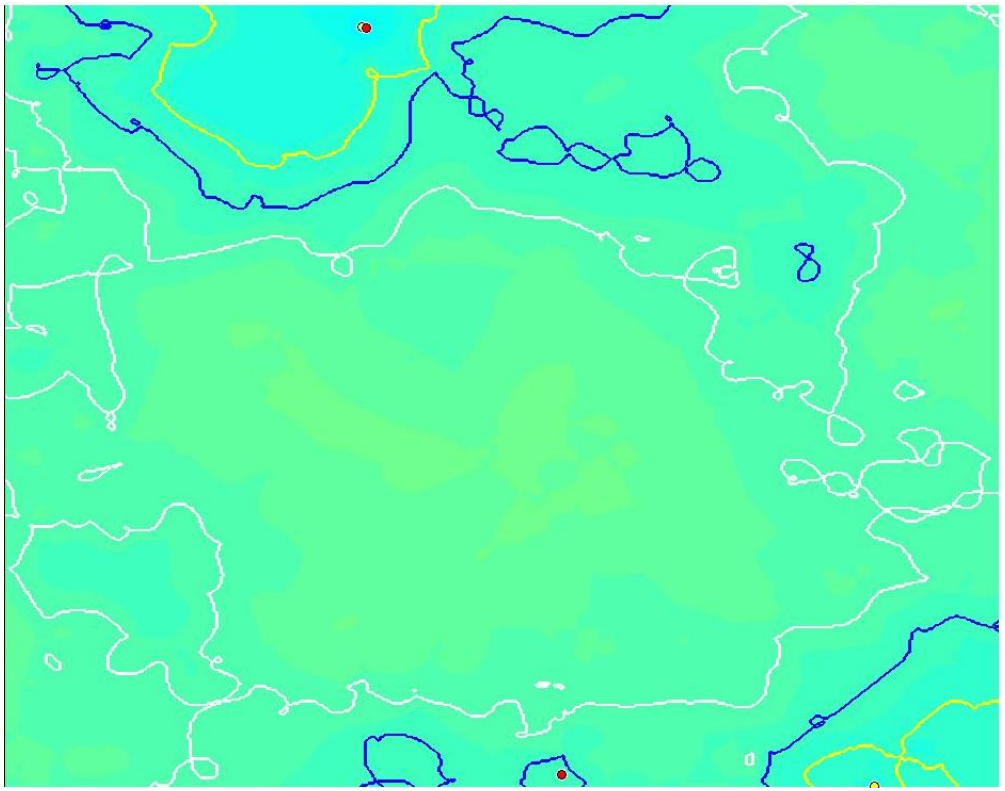
Другой способ:



разумно решать комбинацией двух

Трудности большого числа дыр:





переход к линиям уровня

Главное – выбор эффективной визуализации.

Что часто делается в начале задачи

Задача «Give Me Some Credit»

Статистика признаков

Анализ отдельных признаков: значения должны быть на отрезке [0,1], но есть неожиданные значения + Наны.

Значения

%	Age	#30-59	%	Доход	#o1	#90	#	#60	# в сем
82404 от 0 до 1 потом … Есть дроби!!!	0 (1), 21-109	0-13, 96, 98	70097 до 1 потом… Есть дроби!!!	целые	0-58	0-17, 96, 98	0-26, 32, 54	0-9, 96, 98	0-10, 13, 20

Уникальных значений

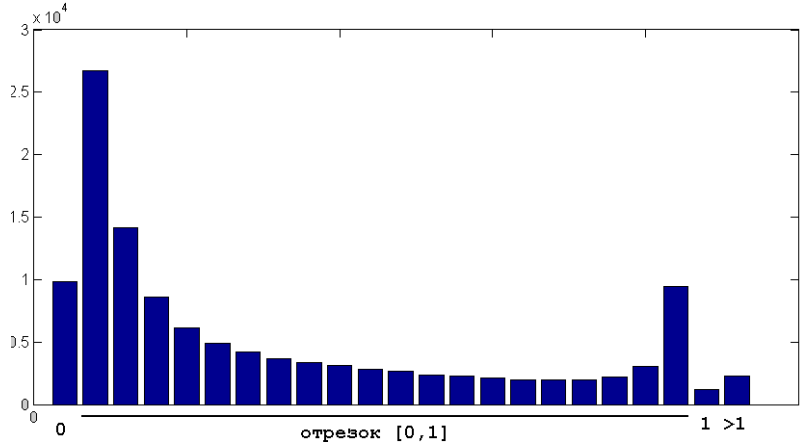
1	2	3	4	5	6	7	8	9	10
84500	86	16	78795	11866	54	19	26	12	13

Нанов

				19831					2630 если тут, то в 5

Аук, Аук через плотность

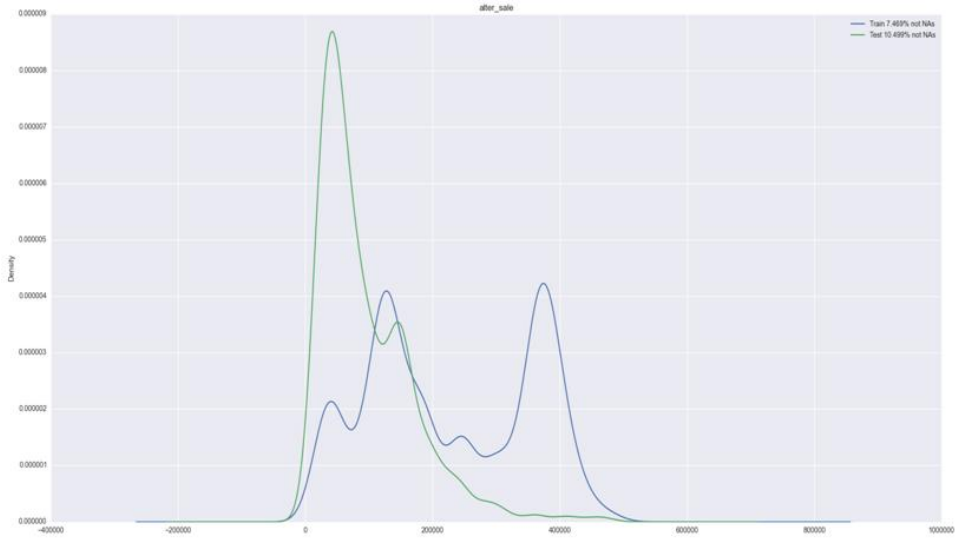
0.7807	- 0.6356	0.6910	0.5266	- 0.5782	- 0.5484	0.6613	- 0.5383	0.6247	0.5482
0.7631	0.6836	0.7077	0.5046	0.6327	0.5518	0.7347	0.5621	0.6525	0.6071
0.7815	0.6329	0.6910	0.5364	0.5554	0.5497	0.6613	0.5432	0.6247	0.5499



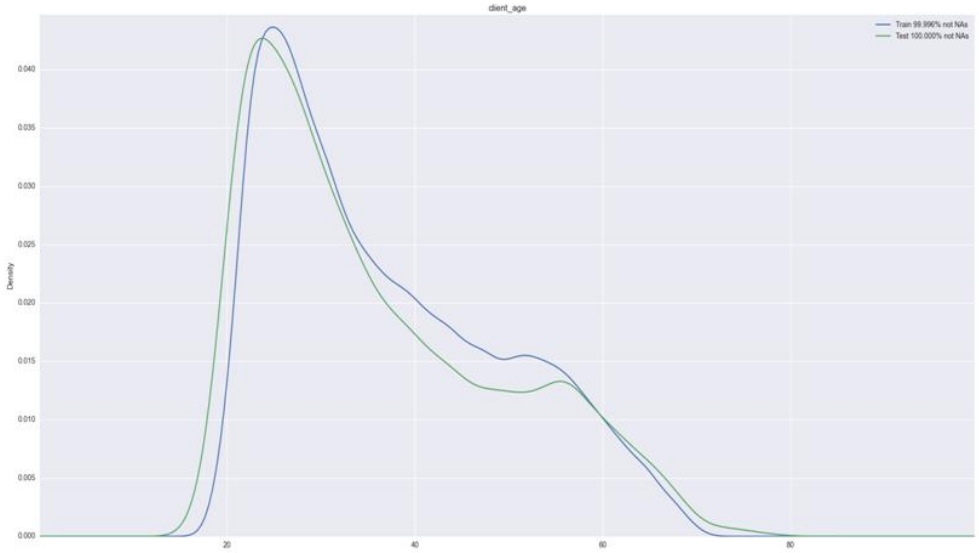
Важно при решении

смотреть как меняются распределения

обучение – контроль



Есть существенные изменения

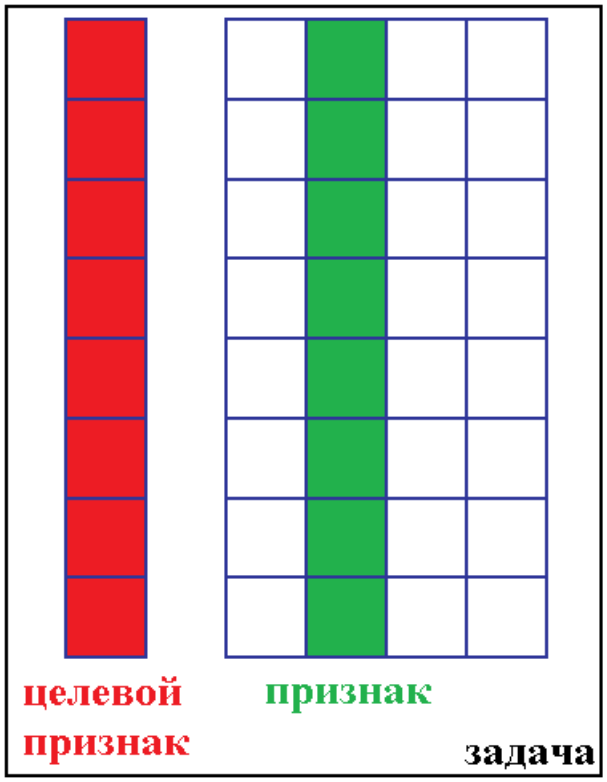


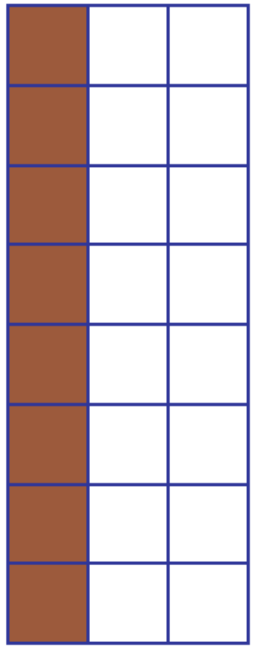
Нет изменений

История про о-трэвел и волшебный признак.

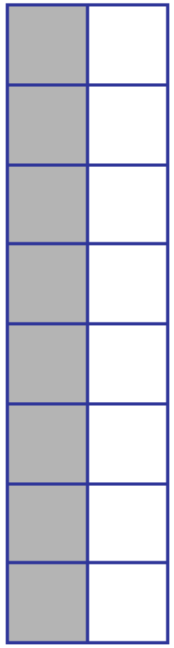
Что можно визуализировать:

«Всё вертикальное»





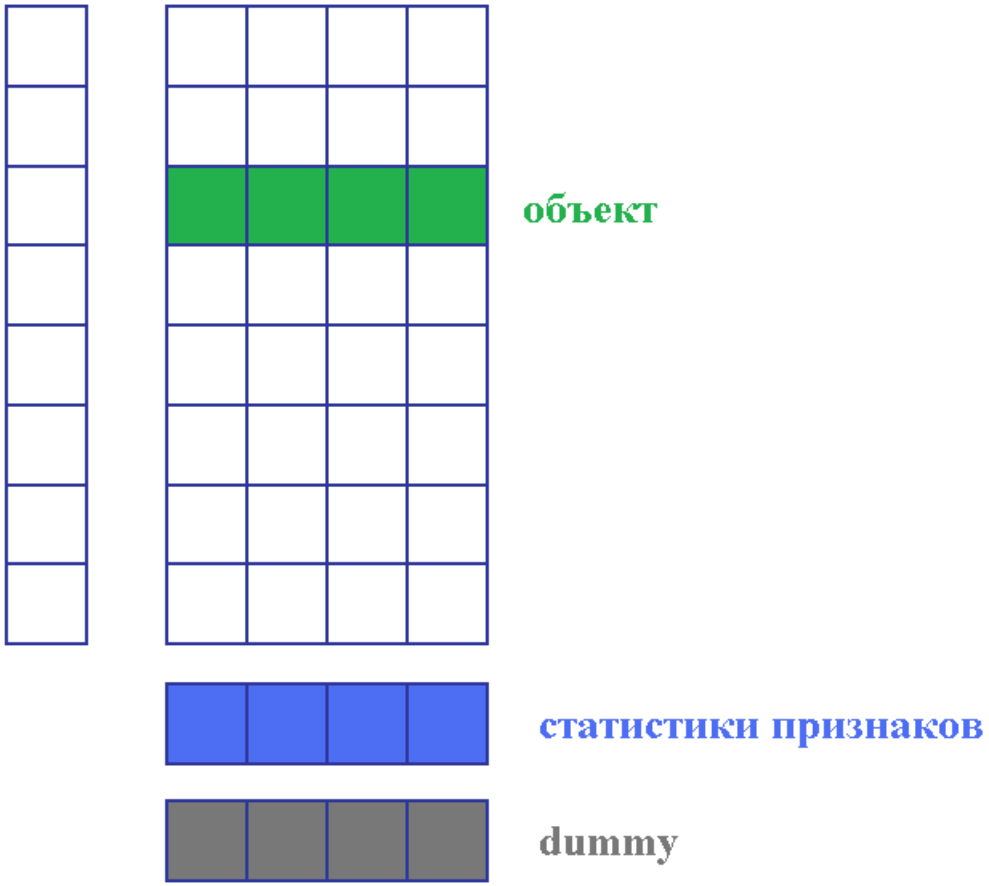
ответы алгоритма



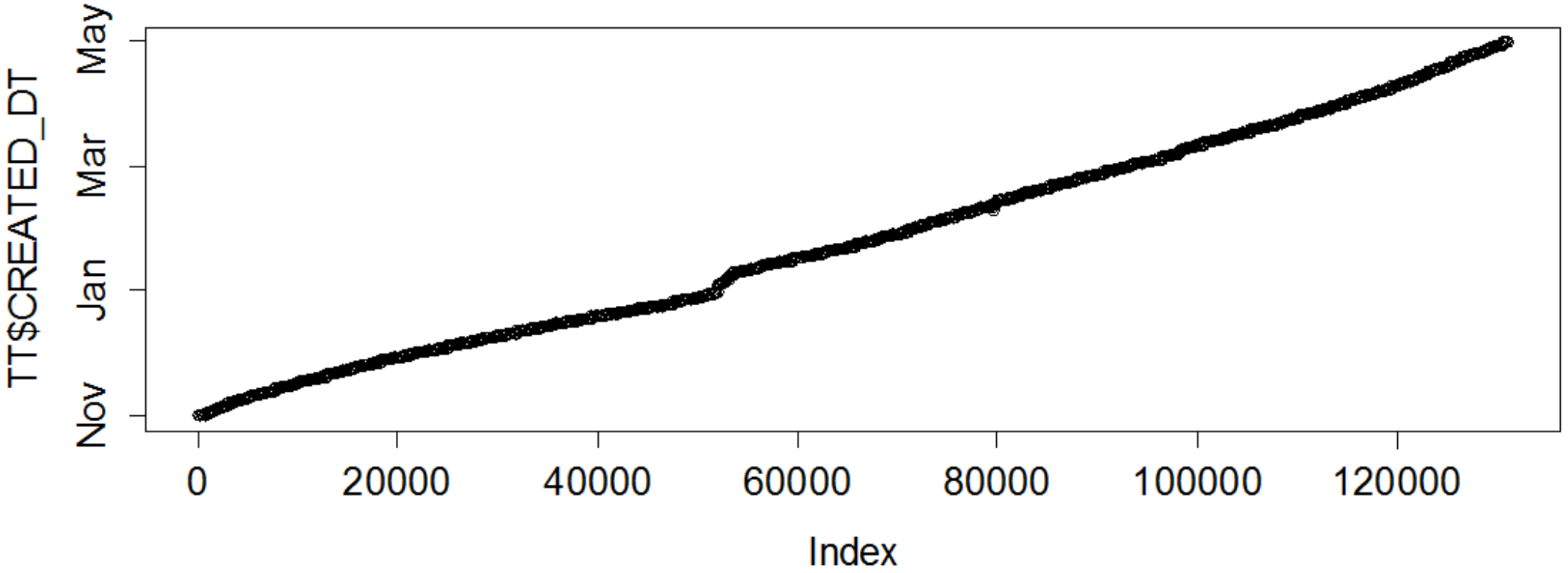
dummy-признаки (номера объектов, групп, rand)

Что можно визуализировать:

«Всё горизонтальное» (реже)



Пример dummy-визуализации

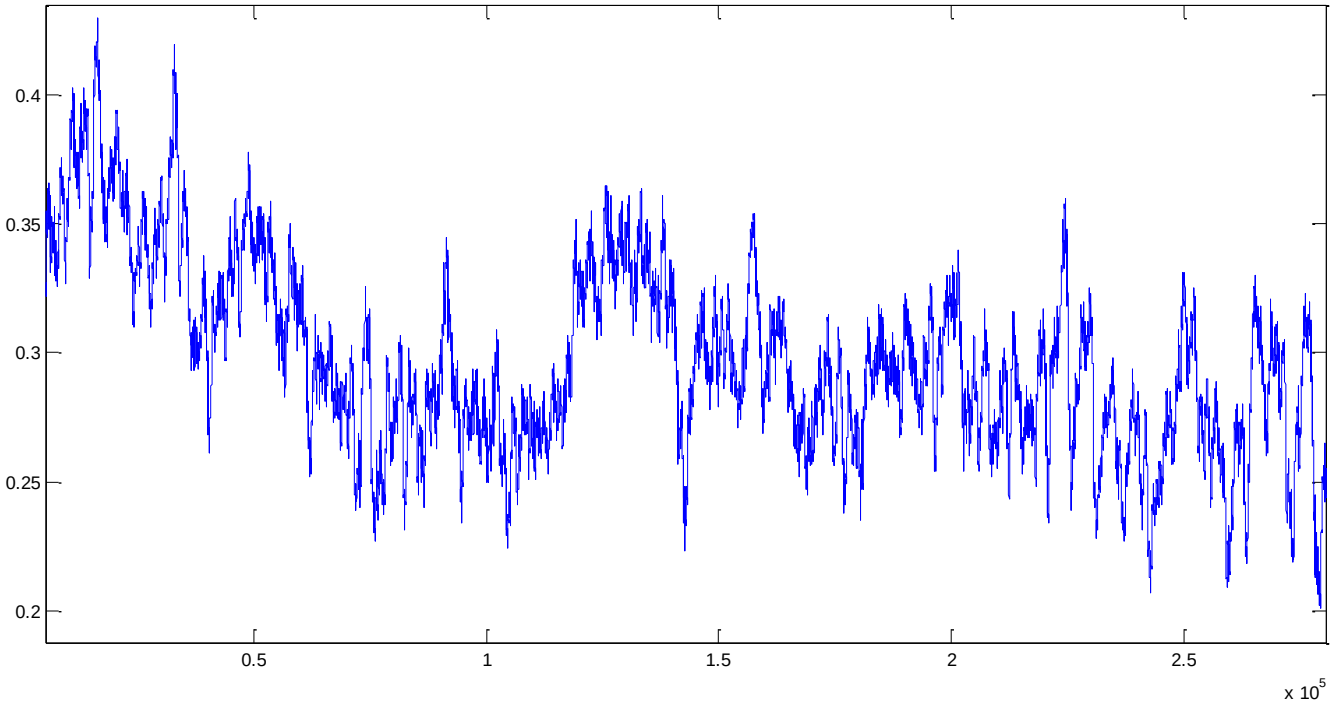


Сделайте график «id – время»:

- простая проверка на монотонность
- видны «подозрительные периоды»

Пример dummy-визуализации

Как меняется цель со временем



Применяется сглаживание окном

Удивительно, но при визуализации:

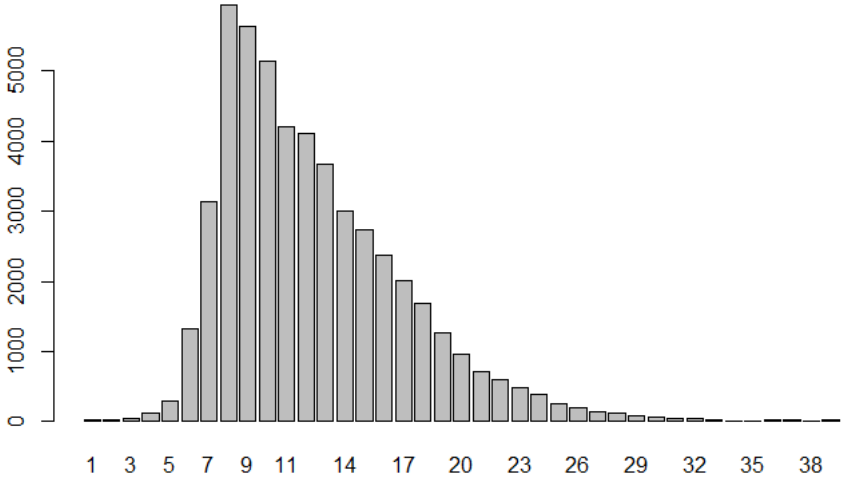
- гладкость
- монотонность или унимодальность
- м.б. + явные выбросы

Если этого нет:

- ищем ошибку

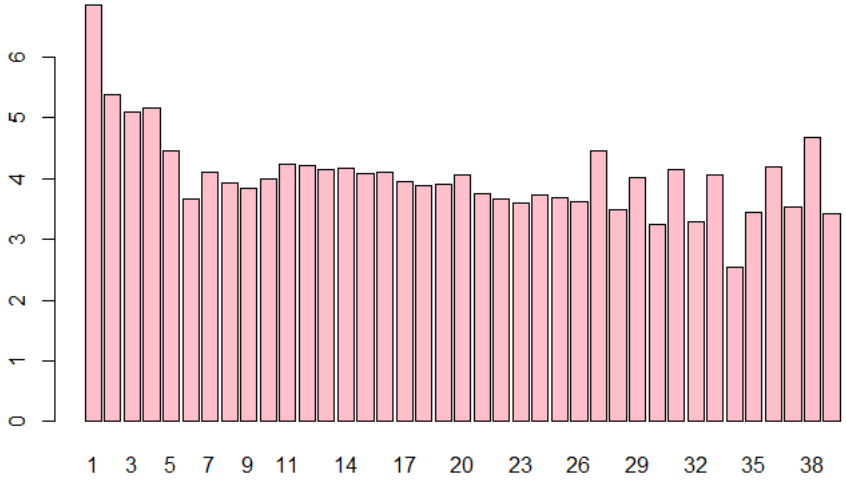
Задача «Liberty»

Целочисленный признак – вещественный или категориальный?




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

Распределение значений признака



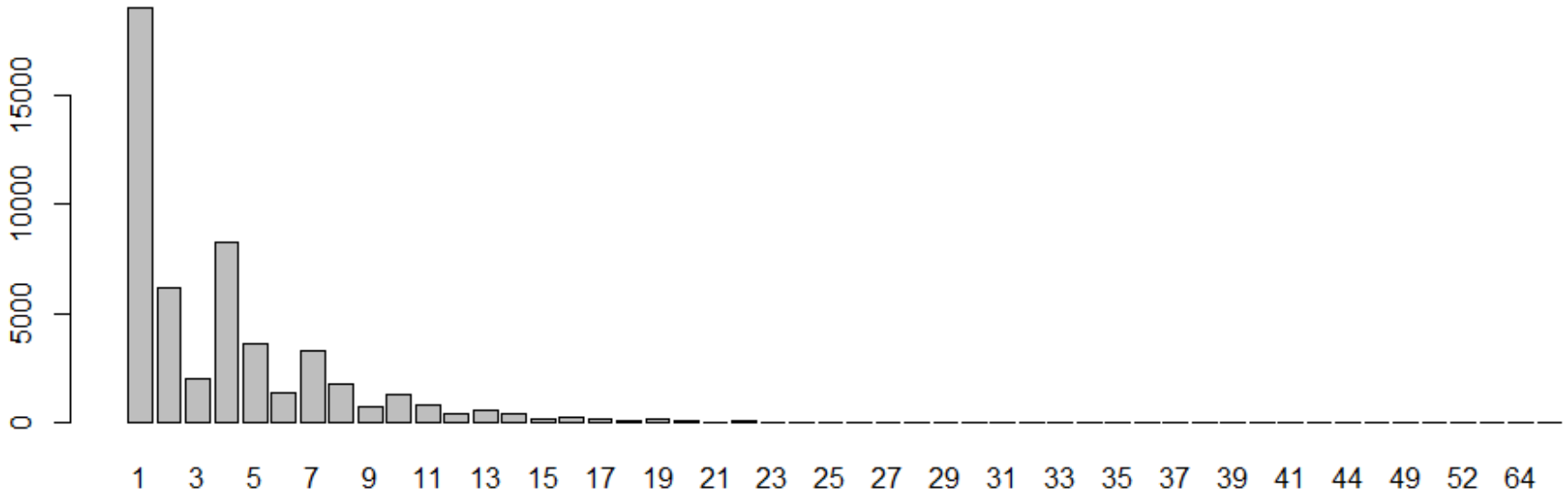

```
barplot(tapply(train$Hazard, train[,34], mean),
        col='pink')
```

Среднее цели на значениях признака

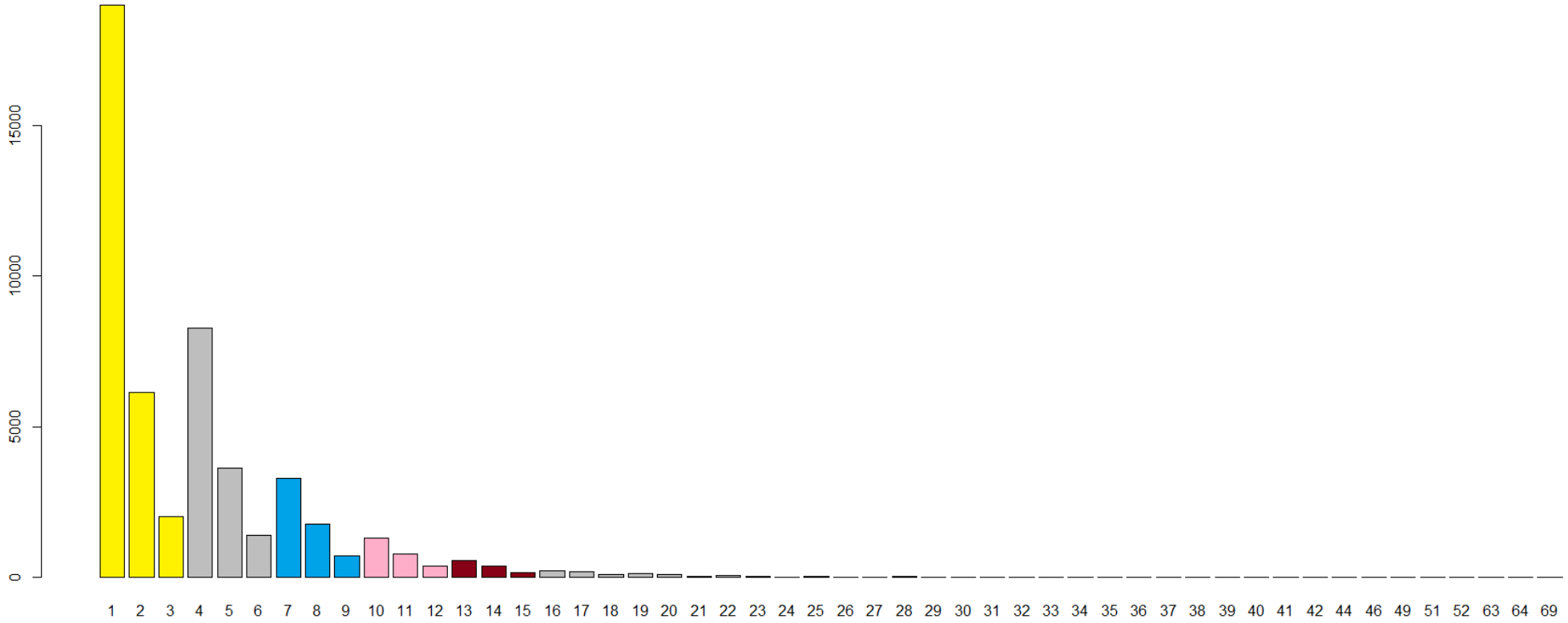
Задача «Liberty»

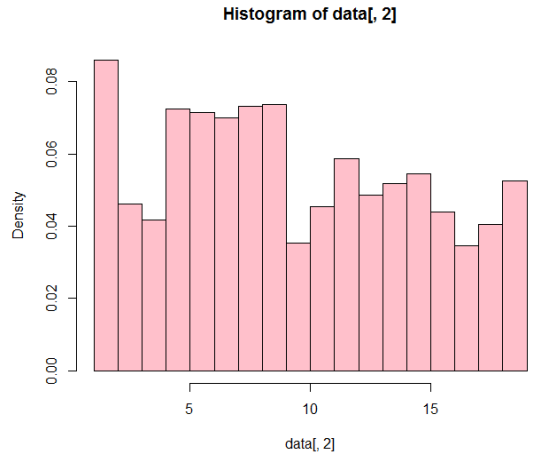
Что интересного в распределении целевого признака?

a transformed count of hazards or pre-existing damages



Задача «Liberty»

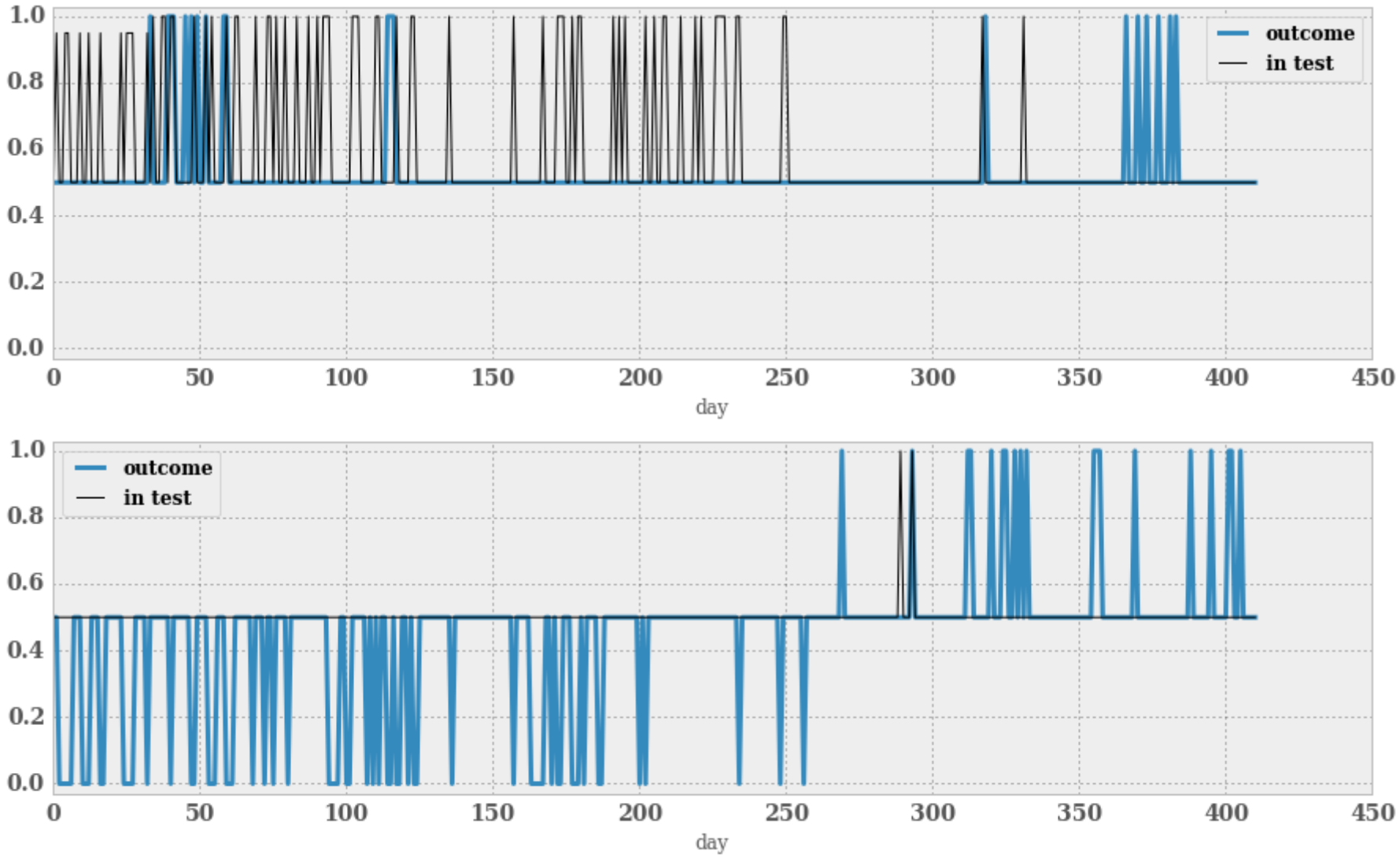




Почему плохо пользоваться стандартными функциями визуализации

```
hist(data[,2], probability = TRUE, col='pink')
```

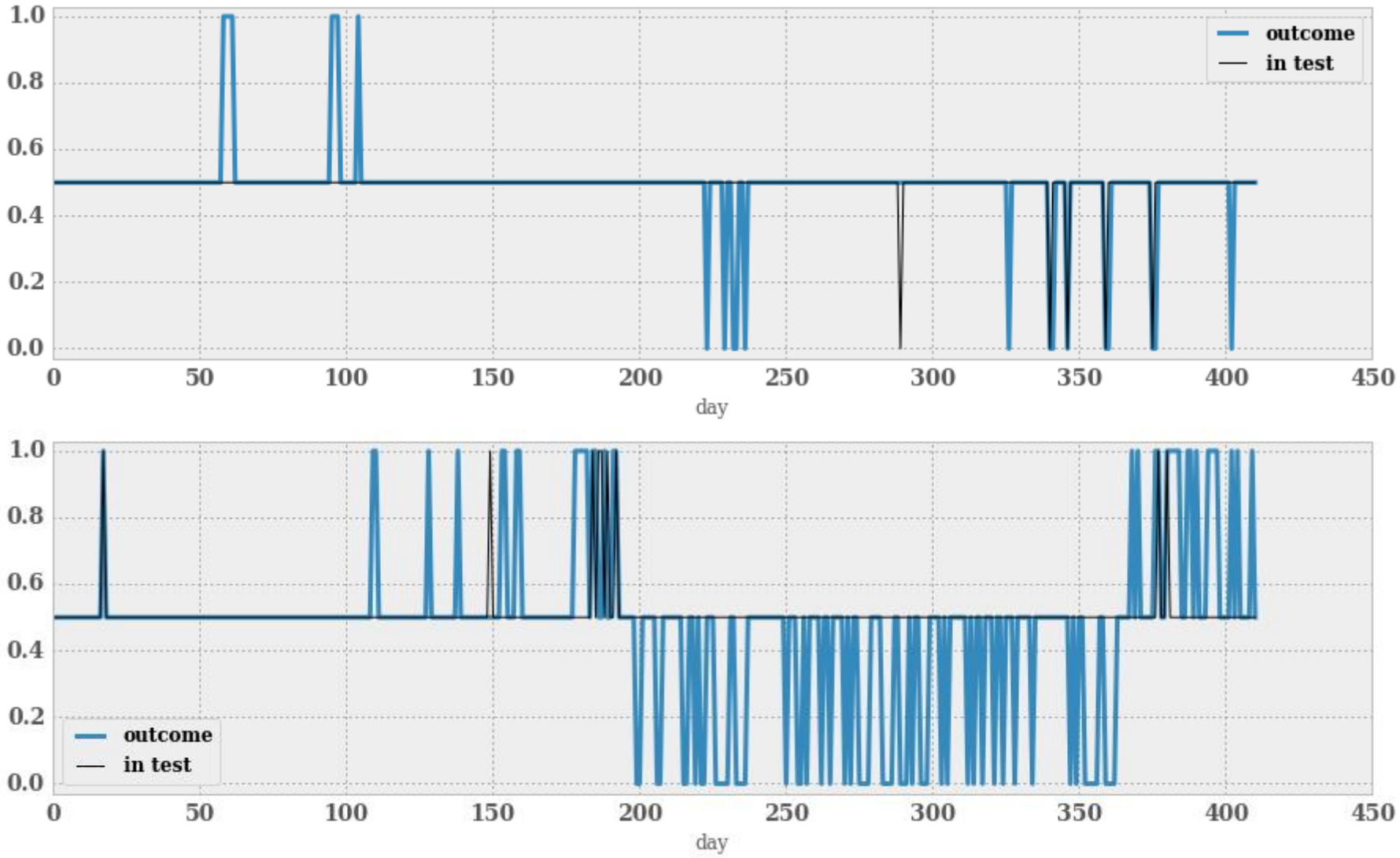
Задача «RedHat»



Как ведут себя представители групп по дням

Каждый график – для отдельной группы

Задача «RedHat»



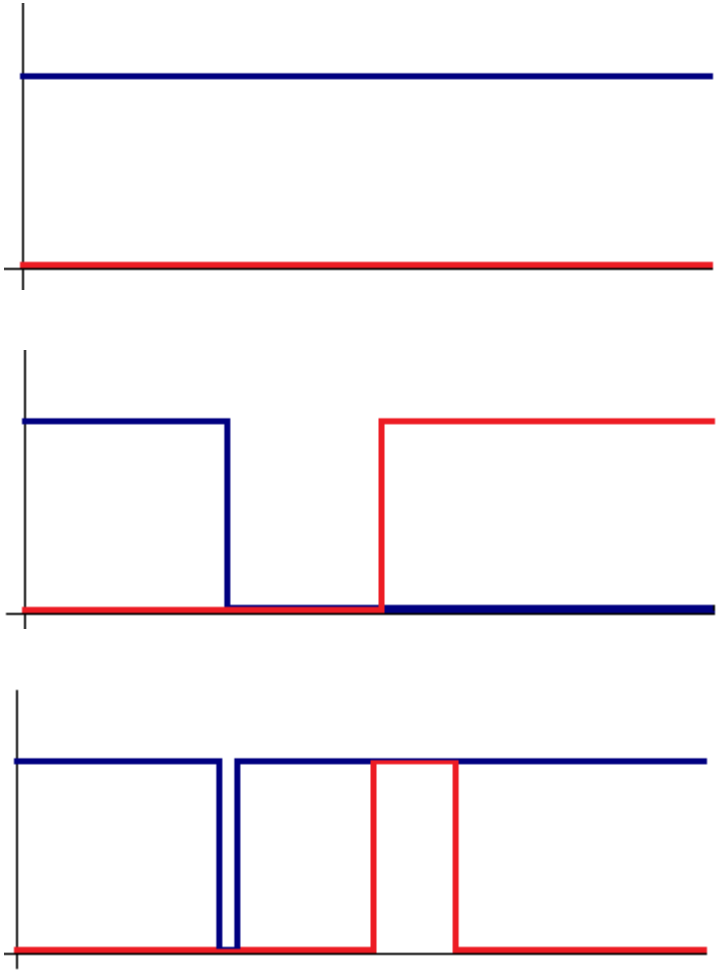
Как ведут себя представители групп по дням

Каждый график – для отдельной группы

Задача «RedHat»

Что видим?

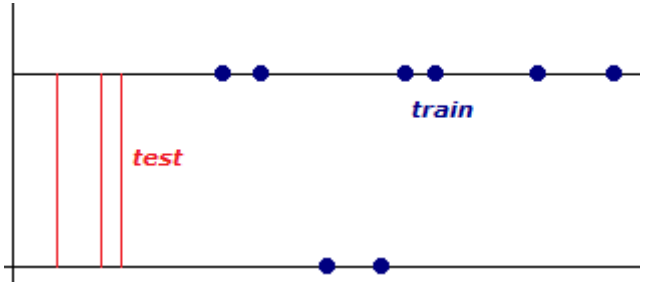
целевой признак кусочно-константный



Причём, максимум 2 «перепада»

Обучение и контроль распределены случайно…

Нет такого…



Задача «RedHat»

Подобные закономерности сложно увидеть в таблице…

	people_id	activity_id	date_x	activity_category	char_1_x	char_2_x	char_3_x	char_4_x	char_5_x	char_6_x	char_7_x	char_8_x	cha
189103	ppl_99966	act2_1740163	2022-09-23	type 2	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.9
189103	ppl_99966	act2_1882139	2022-09-24	type 4	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.9
189103	ppl_99966	act2_3544055	2022-09-27	type 2	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.9
189103	ppl_99966	act2_4300471	2022-09-24	type 2	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.9
189103	ppl_99966	act2_4353827	2022-09-24	type 2	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.9
189103	ppl_99966	act2_4367217	2022-09-23	type 4	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.9
189103	ppl_99966	act2_4459718	2022-09-24	type 4	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.999	-1.9

Так не видно…

Задача «RedHat»

	people_id	date_x	activity_category	outcome
189103	ppl_99966	2022-09-23	type 2	1
189103	ppl_99966	2022-09-24	type 4	0
189103	ppl_99966	2022-09-27	type 2	0
189103	ppl_99966	2022-09-24	type 2	0
189103	ppl_99966	2022-09-24	type 2	0
189103	ppl_99966	2022-09-23	type 4	1
189103	ppl_99966	2022-09-24	type 4	0

убрали лишние столбцы

А так?

Задача «RedHat»

	people_id	date_x	activity_category	outcome
189103	ppl_99966	2022-09-23	type 2	1
189103	ppl_99966	2022-09-23	type 4	1
189103	ppl_99966	2022-09-24	type 4	0
189103	ppl_99966	2022-09-24	type 2	0
189103	ppl_99966	2022-09-24	type 2	0
189103	ppl_99966	2022-09-24	type 4	0
189103	ppl_99966	2022-09-27	type 2	0

сделали сортировку по времени

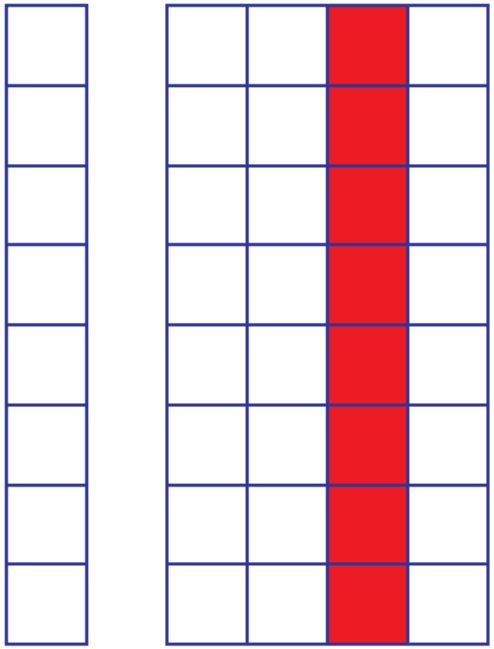
А так?

Полезные операции: группировка и сортировка!

нормировка и tiedrank

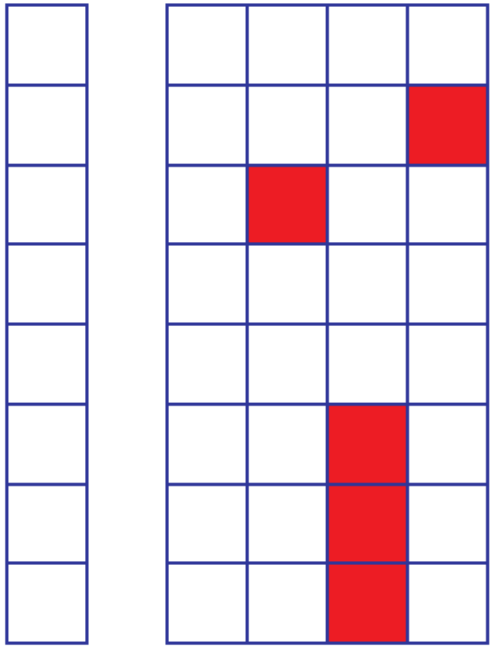
Что есть в данных:

- шумовые признаки



удалить

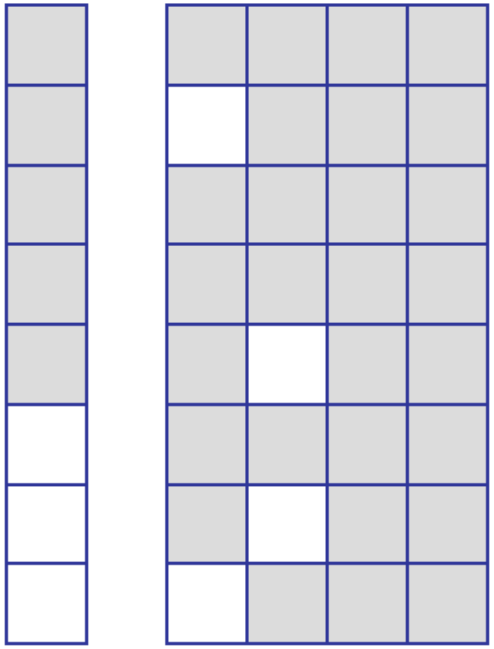
- шумовые значения



причины:

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

-пропуски:



причины:

«нет значения», «не знаем значения»

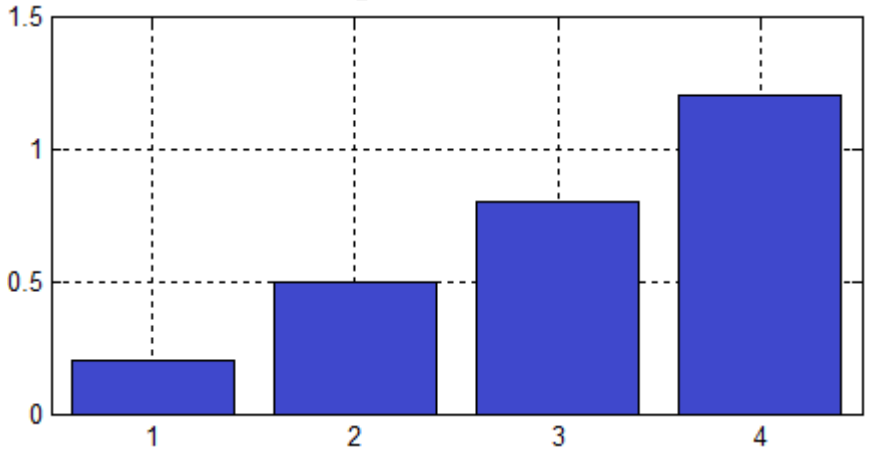
метод:

+dummy!!!

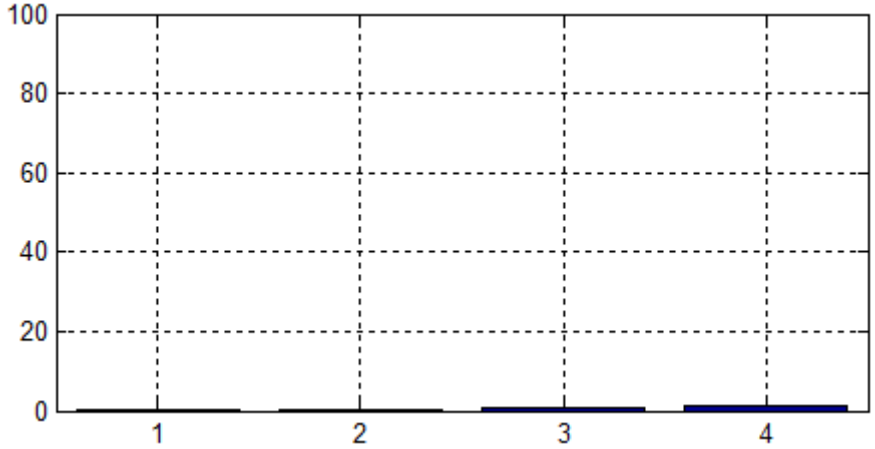
Про рекомендации к визуализации

Процент женщин в парламенте

«неправильно»



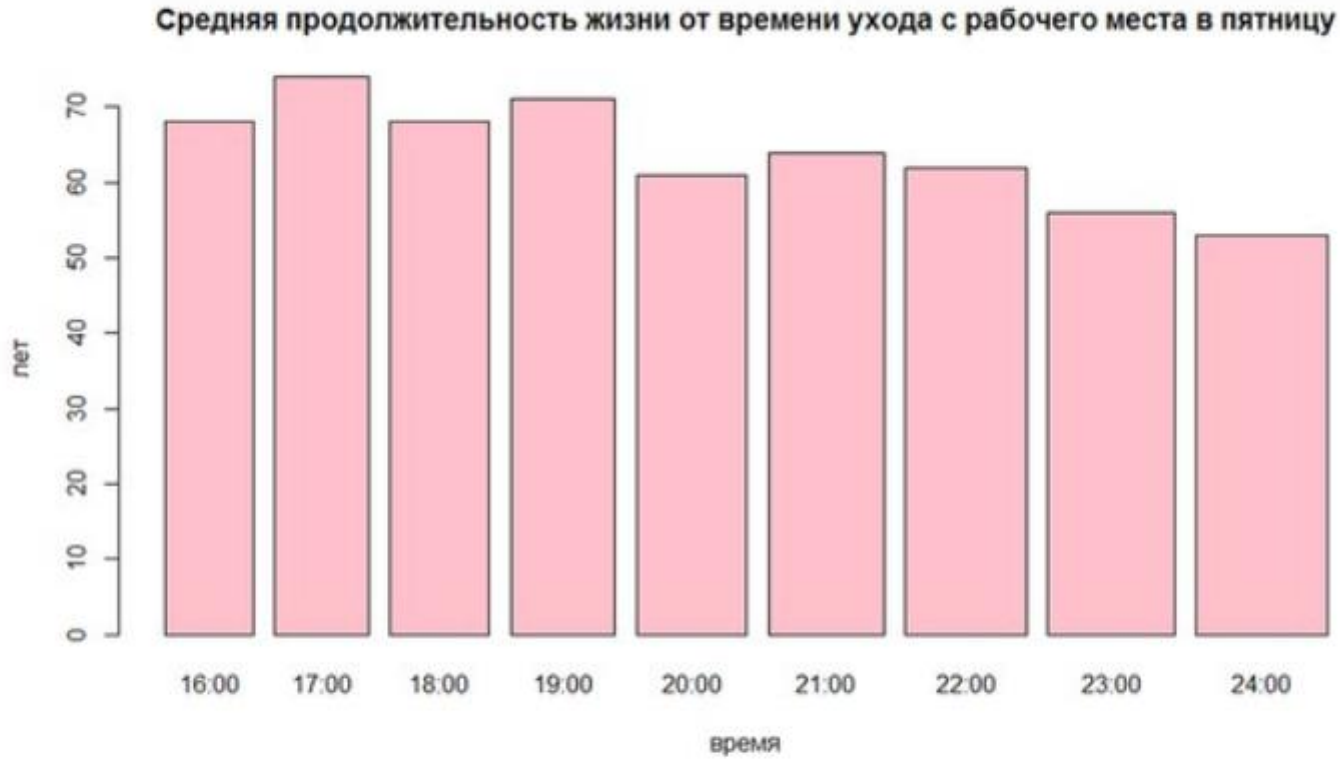
«правильно»



А если это процент убитых в Битцевском парке?

Про рекомендации к визуализации

АлгоМост



24 июл в 12:25 Поделиться Мне нравится 8

масштаб отвратительный

24 июл в 12:43 | Ответить