курс «Прикладные задачи анализа данных»

Искусство визуализации

Часть 4. Кейсы

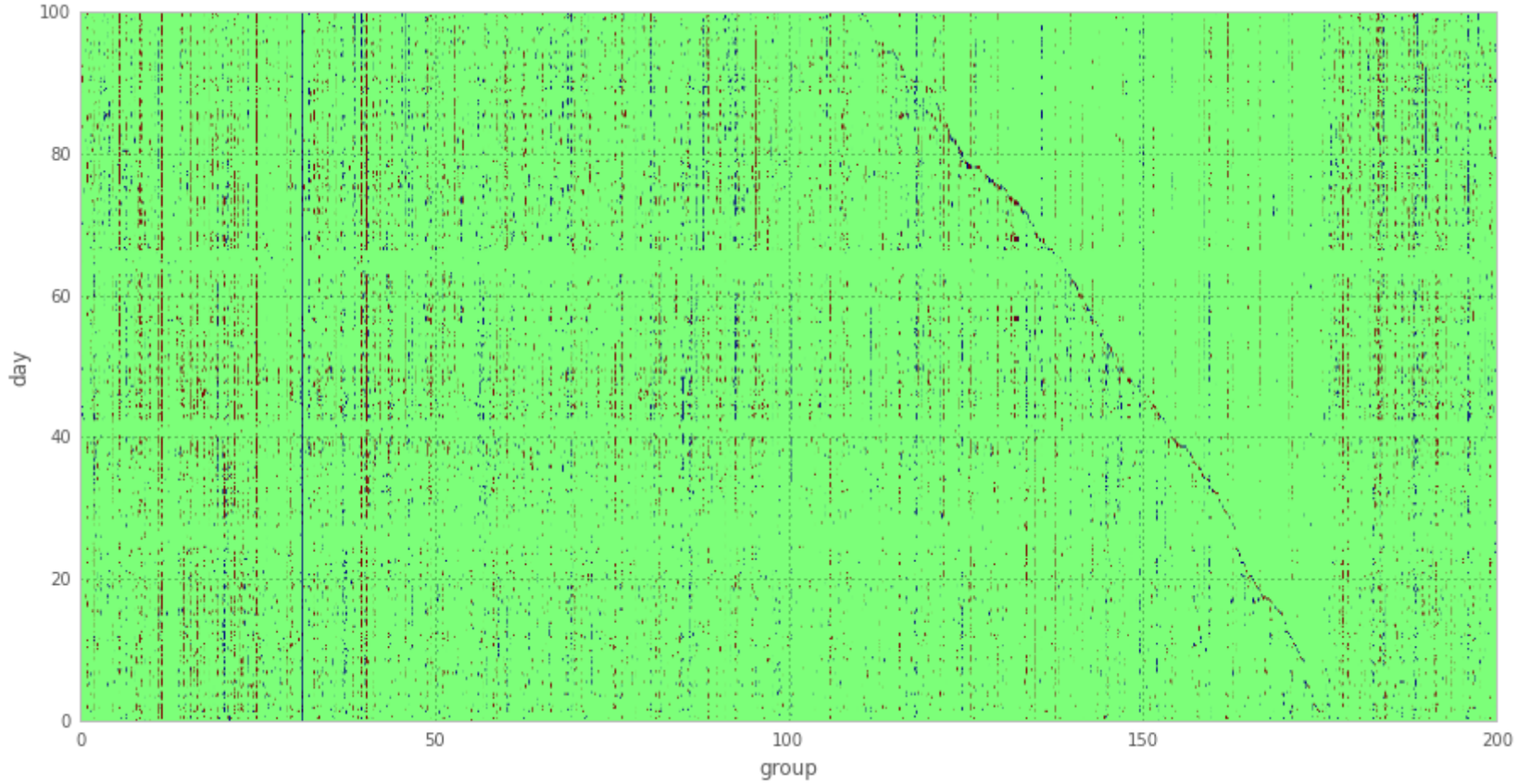
Александр Дьяконов

16 сентября 2019 года

План

- **RedHat:** нумерация групп и закономерности в целевом признаке
- **оценка эффективности менеджера:** изменение распределений признаков во времени
- **«причина-следствие»:** операции над признаками
- **чёрные дыры:** удобные визуализации
- **check-in:** привязка к географии
- **Ascott Group:** группы временных рядов
- **Ticketland ML Contest:** просто иллюстрация результатов поиска
- **Ozon:** бесполезные данные
- **Сбербанк:** гендерные закономерности

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:

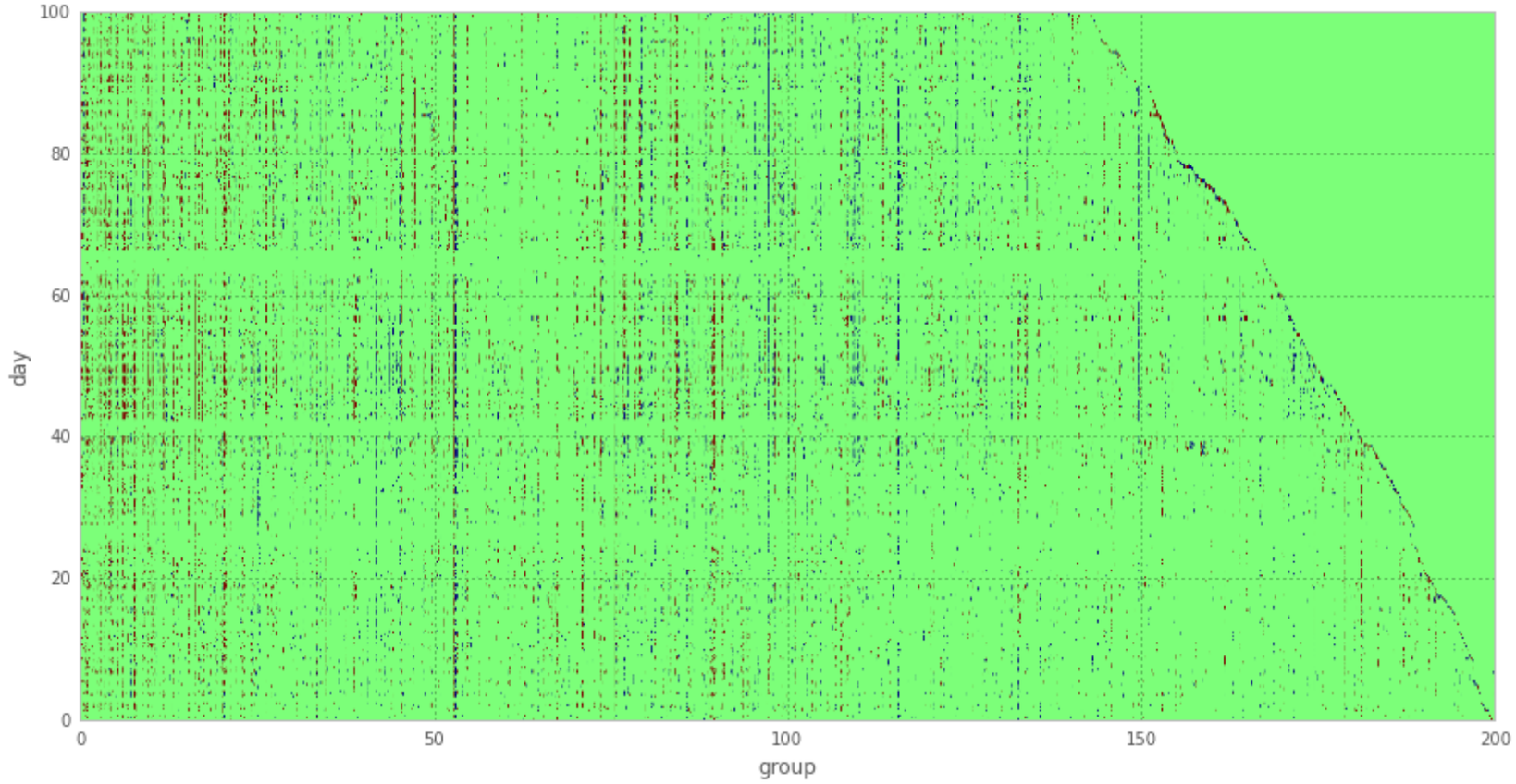
```
group_date2.columns[:10]

'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…

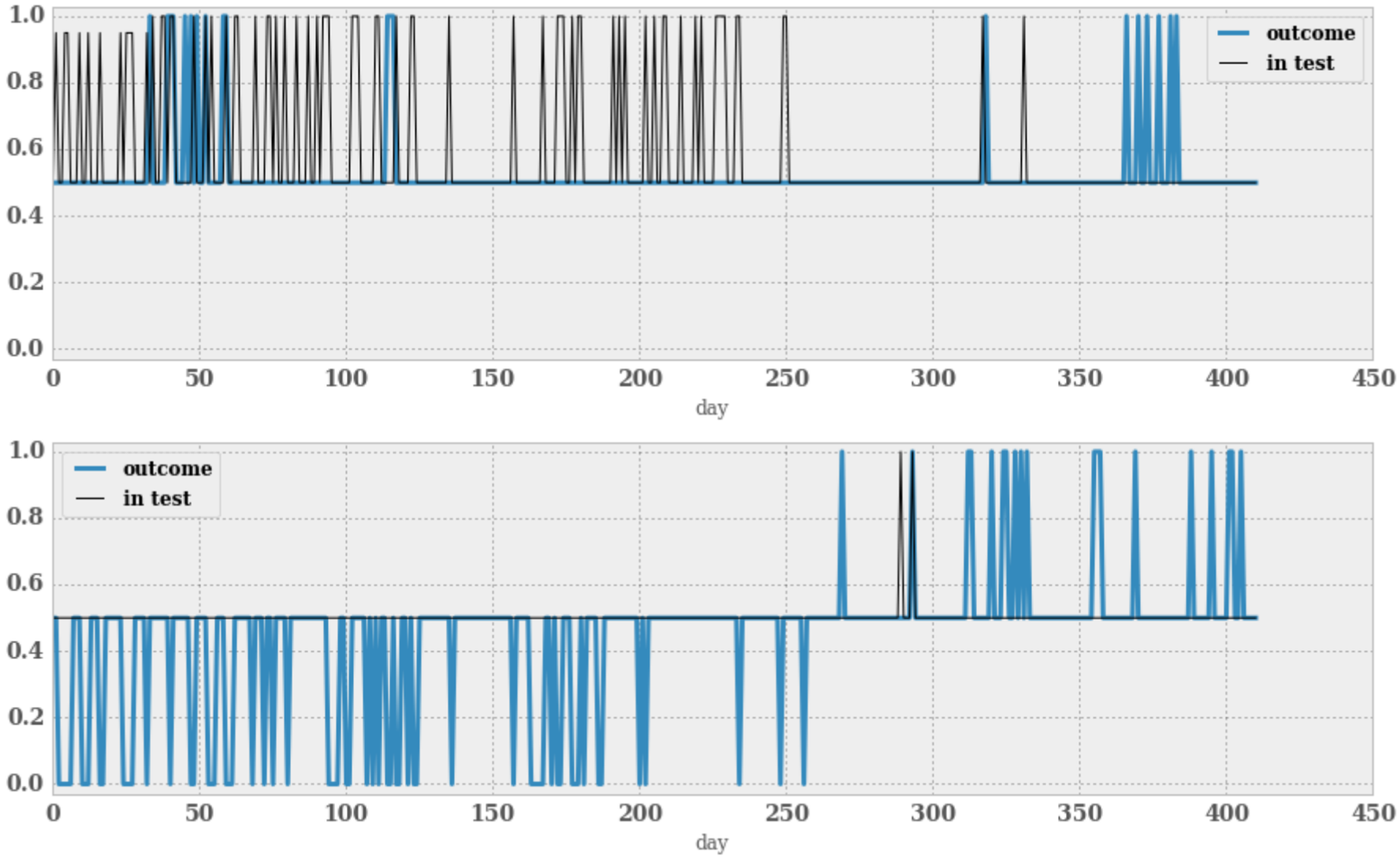
```
data_train.group_1 = data_train.group_1.map(lambda x: int(x[6:]))
```

Визуализация данных – RedHat



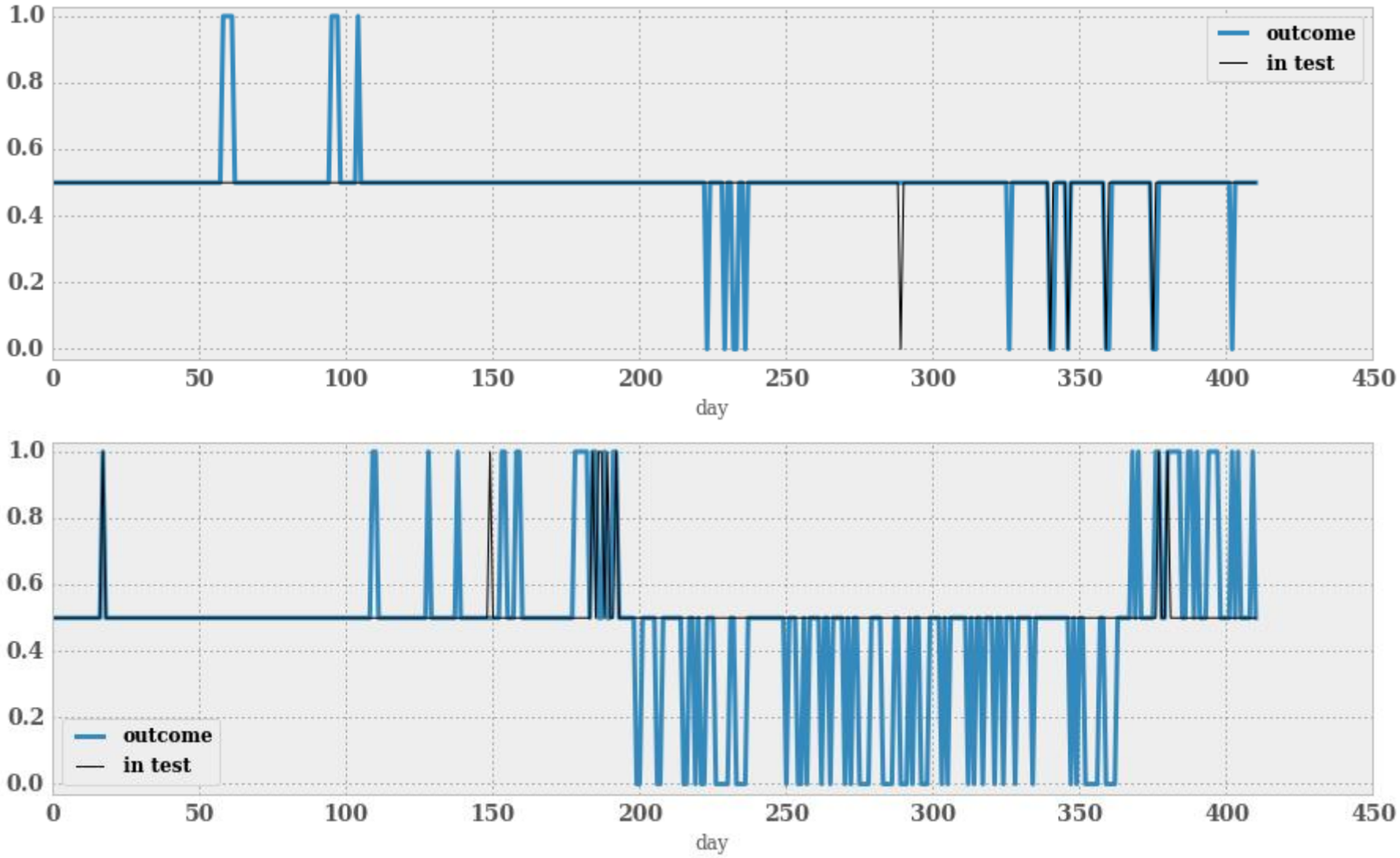
теперь понятнее… группы, видимо, идут в порядке появления
последние – которые добавлялись в дни сбора выборки

Визуализация данных – RedHat



Каждый график – отдельная группа: как ведут себя её представители

Визуализация данных – RedHat

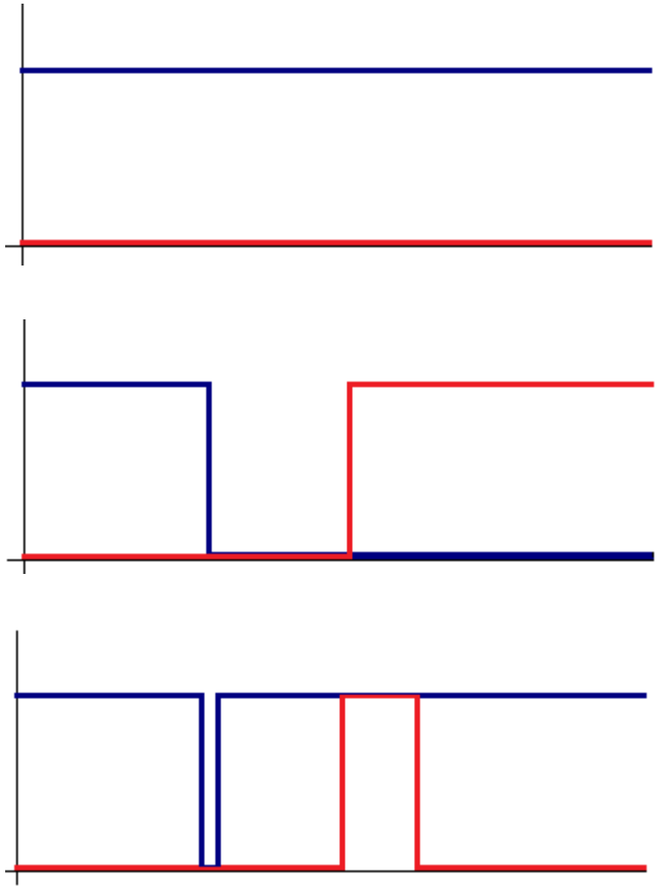


Каждый график – отдельная группа: как ведут себя её представители

что видим?

Визуализация данных – RedHat

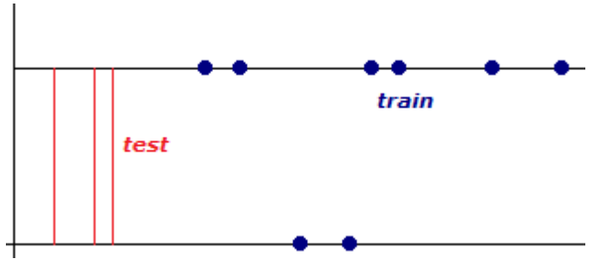
целевой признак кусочно-константный



Причём, максимум 2 «перепада»

Обучение и контроль распределены случайно…

Нет такого…



Визуализация данных – 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

Визуализация данных – 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?

Визуализация данных – 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

Визуализация данных – оценка эффективности менеджера

Дано: описание менеджера и клиента

Целевой признак: Была ли между ними успешная сделка

В обучении: ~9500 Записей, ~22 признака

В тесте: ~4000 записей

Важно: обучение/тест разбиты по времени

Важно: почти все признаки не вещественные (время, факторы)

Функционал качества: AUC ROC

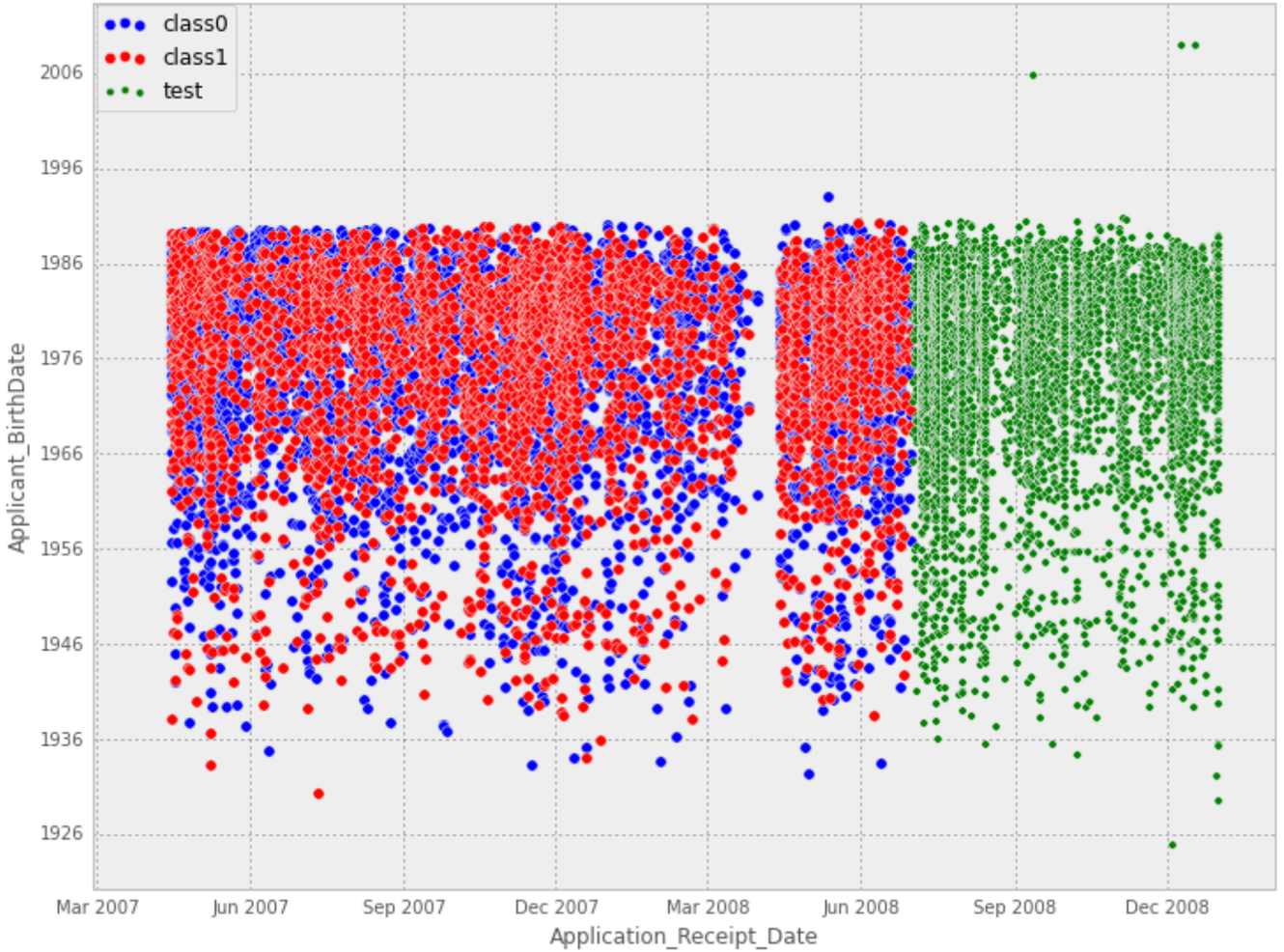
Визуализация данных – оценка эффективности менеджера

Смотрим данные – делаем гипотезы

	ID	Office_PIN	Application_Receipt_Date	Applicant_City_PIN	Applicant_Gender	Applicant_BirthDate	Applicant_Marital_Status
0	FIN1000001	842001	2007-04-16	844120	M	1971-12-19	M
1	FIN1000002	842001	2007-04-16	844111	M	1983-02-17	S
2	FIN1000003	800001	2007-04-16	844101	M	1966-01-16	M

- есть благоприятные дни для сделки?
- на сделку влияют пол менеджера/клиента?
  - посмотреть их разницу в возрасте
- посмотреть успешность/загруженность/опыт менеджера

Визуализация данных – оценка эффективности менеджера



Признак «время сделки» по горизонтали

Что интересно?

Визуализация данных – оценка эффективности менеджера

Если делать контроль CV – качество 0.65 AUC ROC

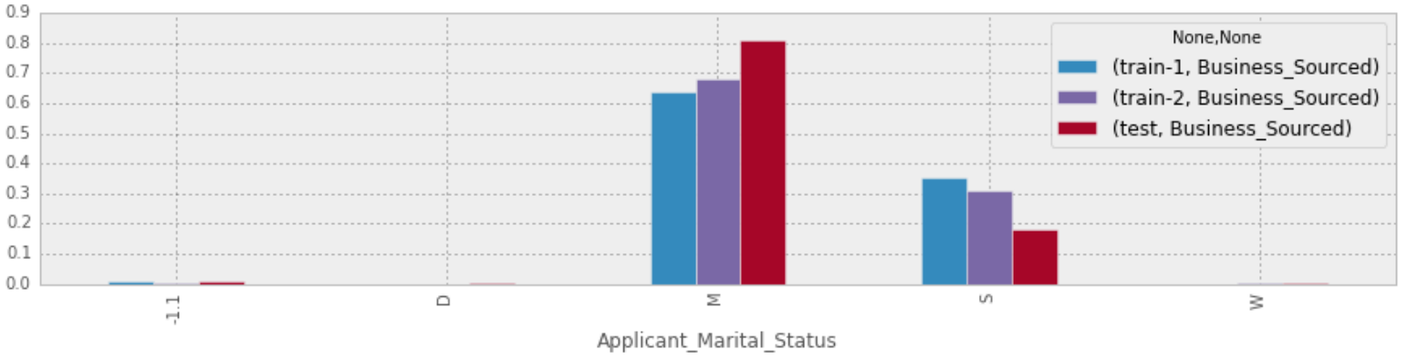
Если контроль – последний кусок обучения – 0.55 AUC ROC

Теперь ясно почему!

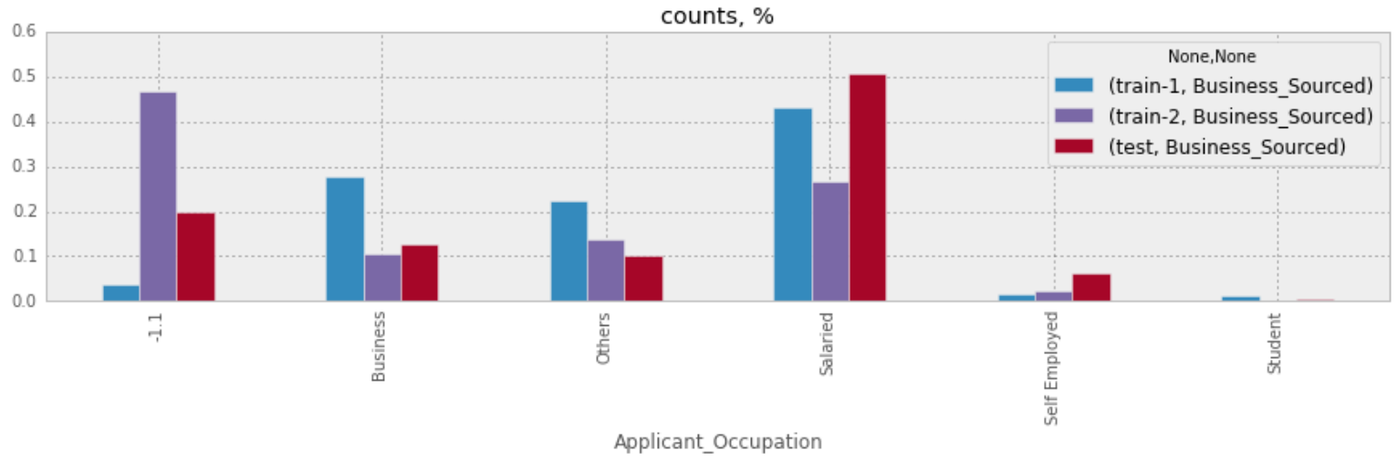
Распределение в разных кусках

Визуализация данных – оценка эффективности менеджера

пол



семейное положение



род занятий

Визуализация данных – оценка эффективности менеджера

Изменение распределений признаков во времени (сделан jitter)

```
{nan:0, 'Self Employed':1, 'Business':2, 'Salaried':3, 'Others':4, 'Student':5}
```

Визуализация данных – оценка эффективности менеджера

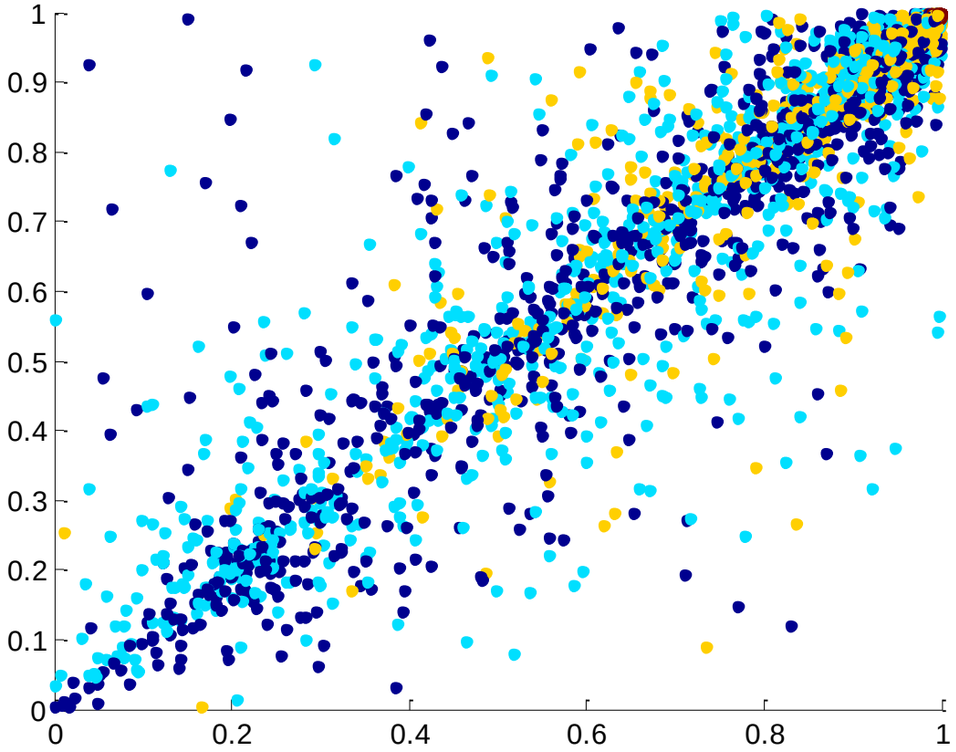
Интересный приём:

по train1 кодировать признаки,

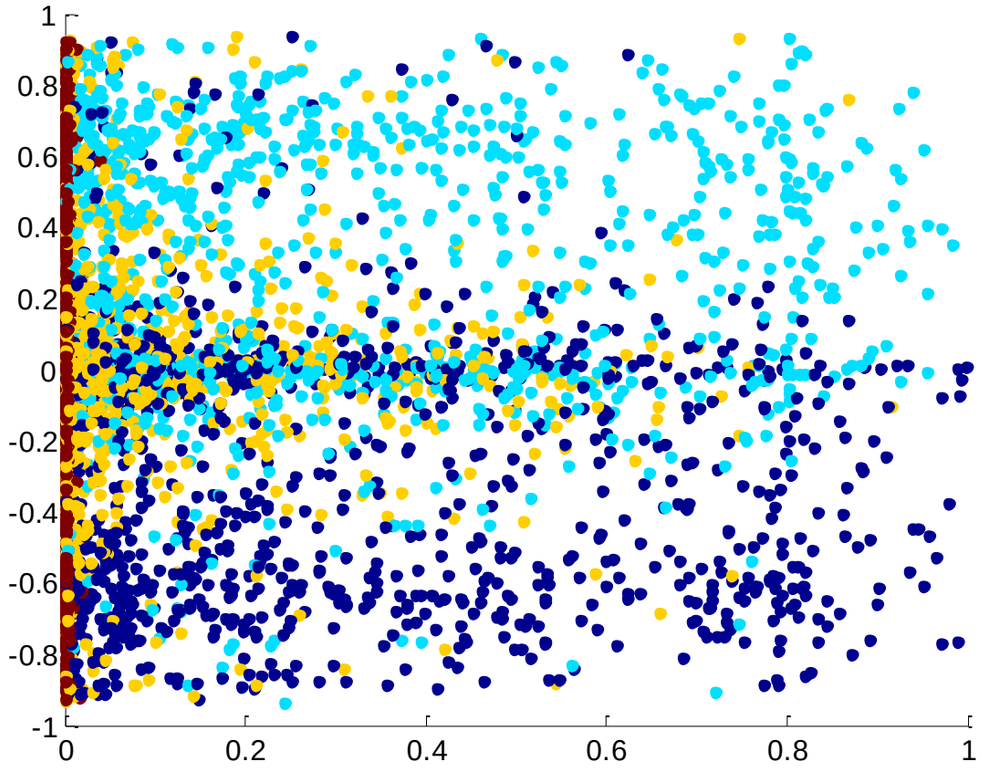
на train2 обучать...

Визуализация данных – задача «причина-следствие»

Метод: «ручная деформация пространств»



f, g

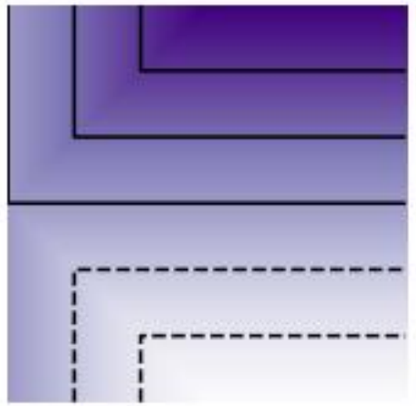


1-0.5*(f + g), fn - gn

алгебраические выражения над признаками

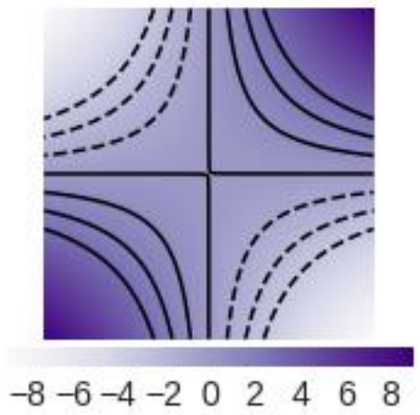
Визуализация данных – задача «причина-следствие»

А теперь надо «уголками откусывать классы»:

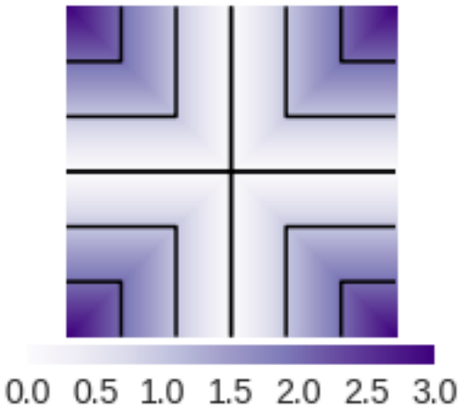


Какие функции «откусывают уголками»

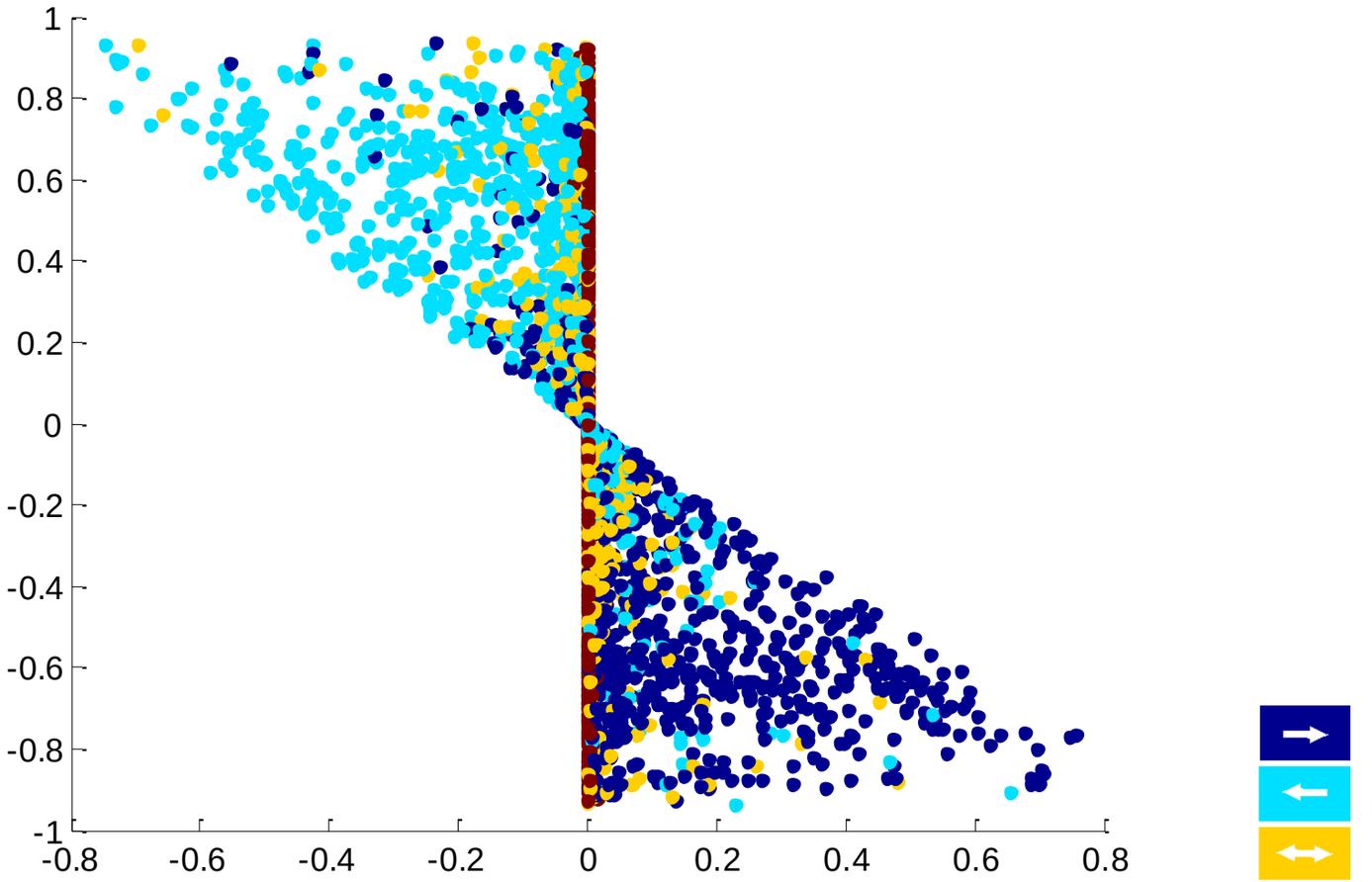
$$z = y \cdot x$$



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



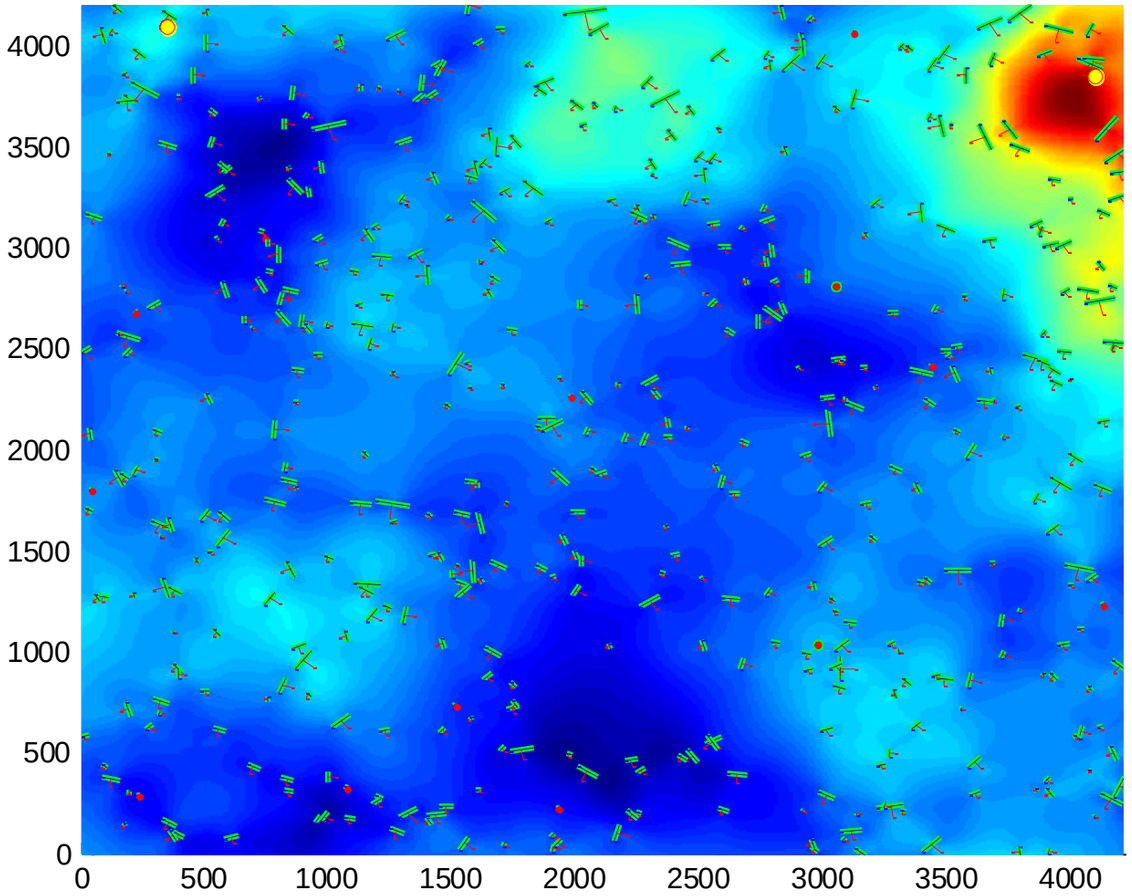
Визуализация данных – задача «причина-следствие»

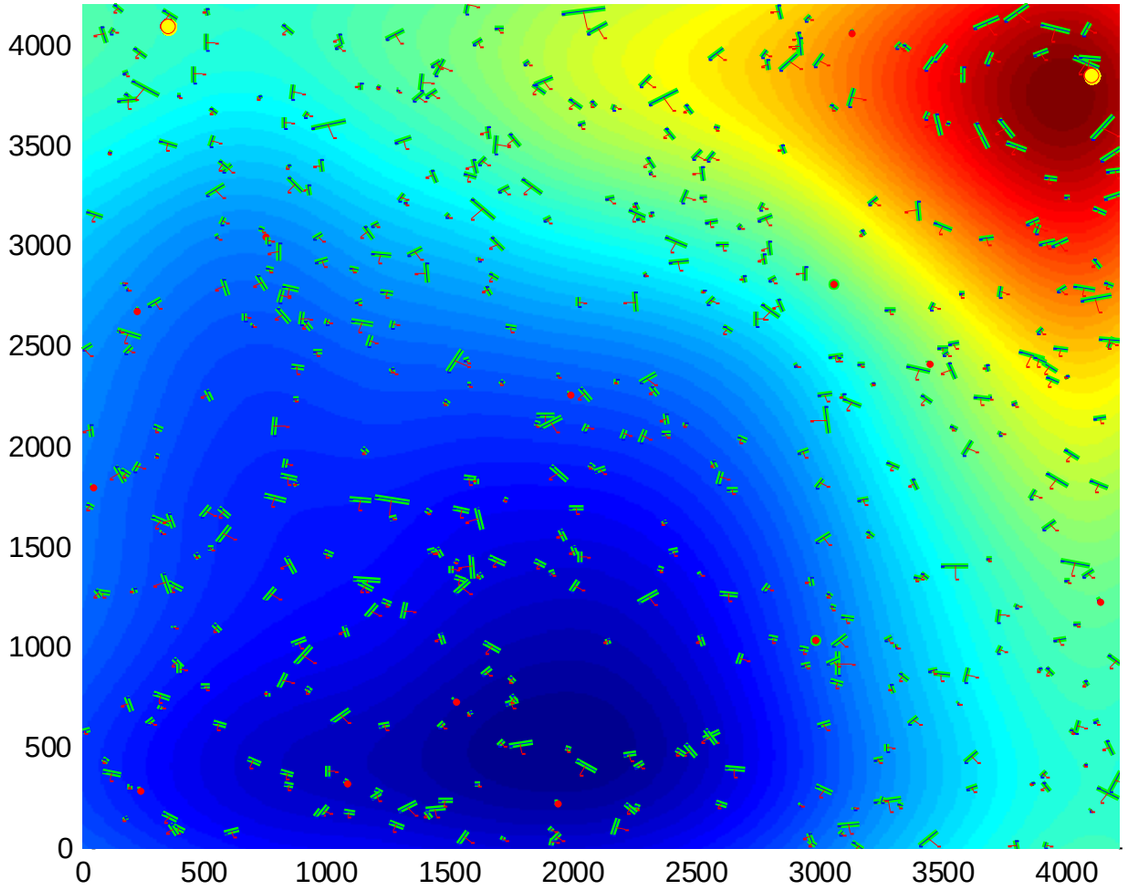



```
-(1-0.5*(f+g)).*(fn-gn)
```

И здесь мы видим разделяемость синих и голубых!

Визуализация данных – задача про чёрные дыры



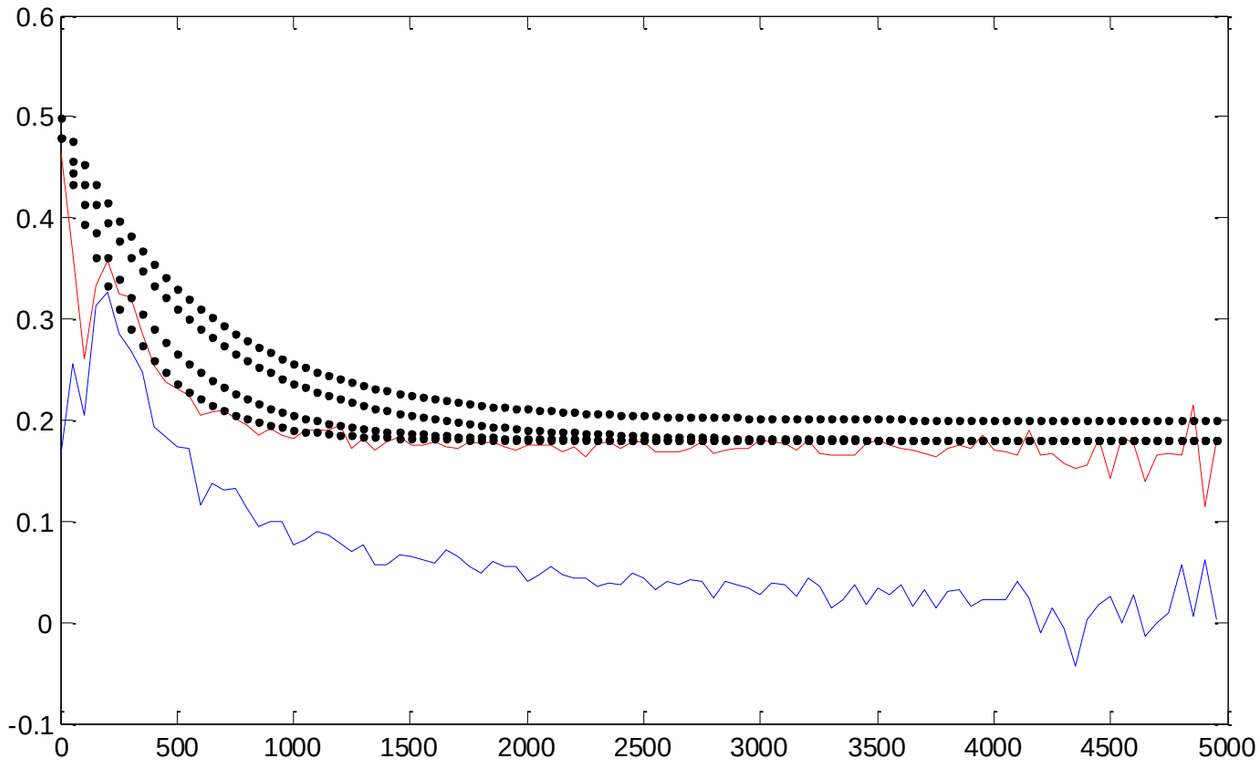


Какая связь между рисунками?

Визуализация данных – задача про чёрные дыры

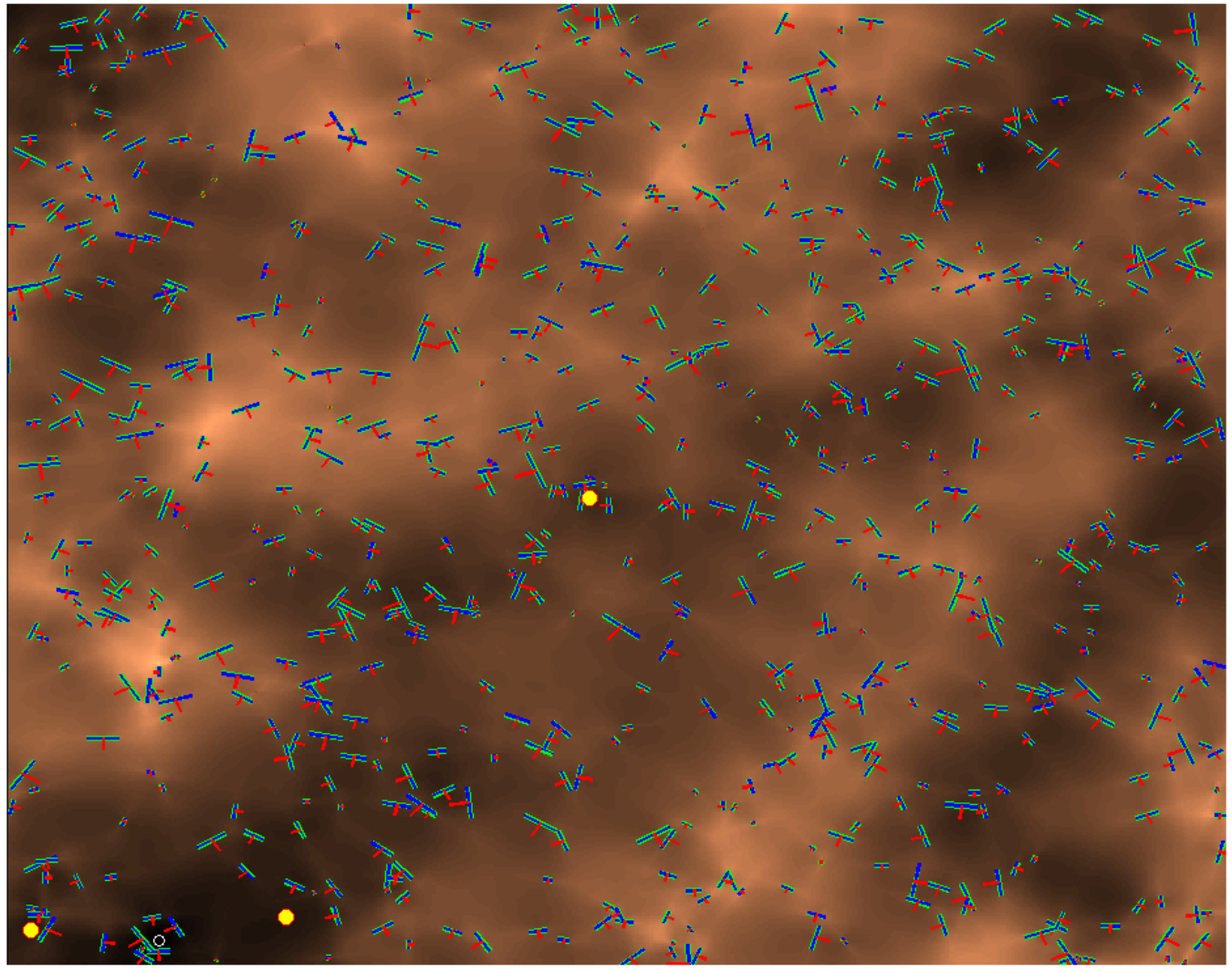
Ответ: «Плотность» и её сглаженный аналог.

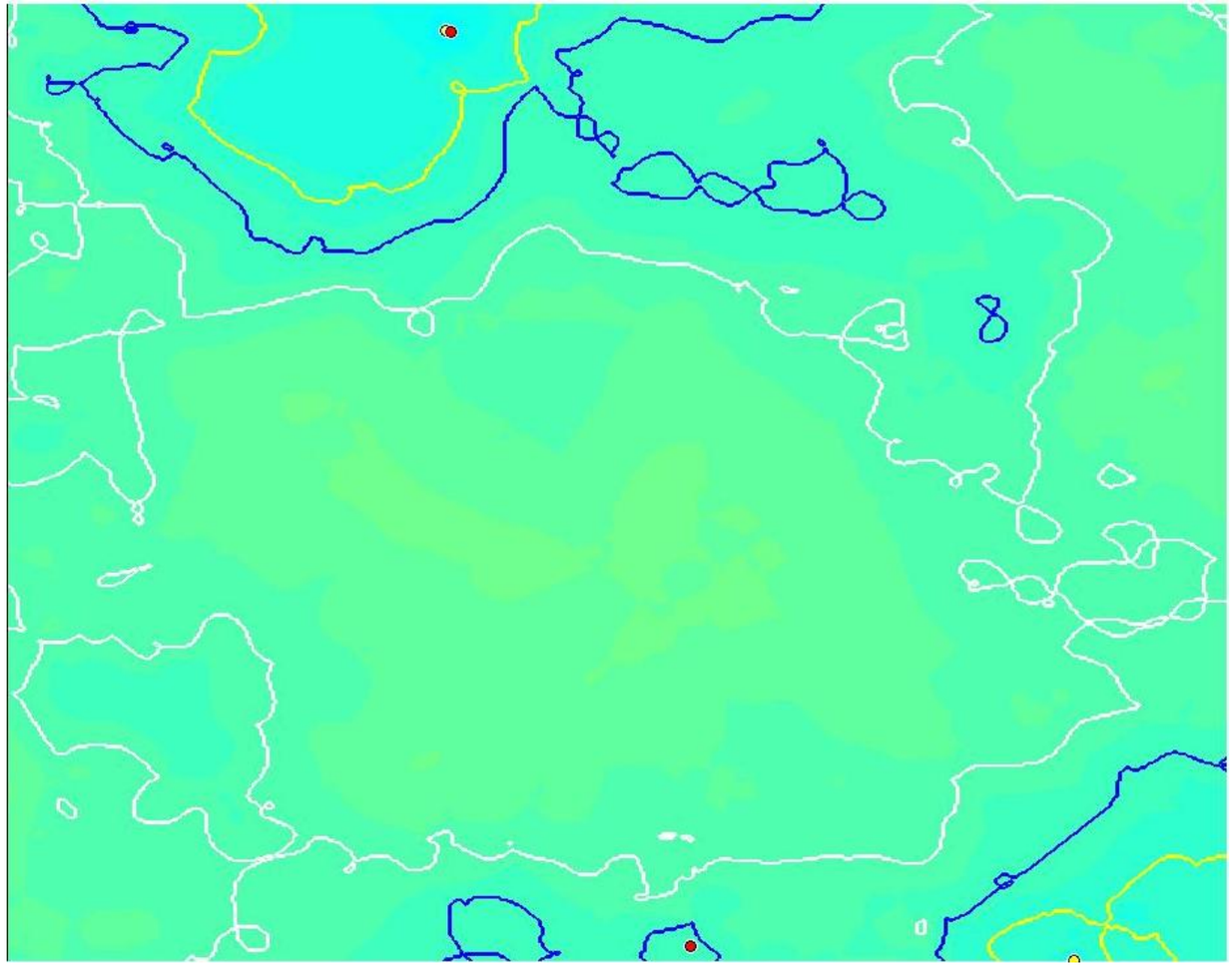
Средний профиль плотности(красный):



и методы его приближения

Визуализация данных – задача про чёрные дыры





Трудности большого числа дыр

Главное – выбор эффективной визуализации – переход к линиям уровня

Визуализация данных – по какому принципу упорядочены данные?

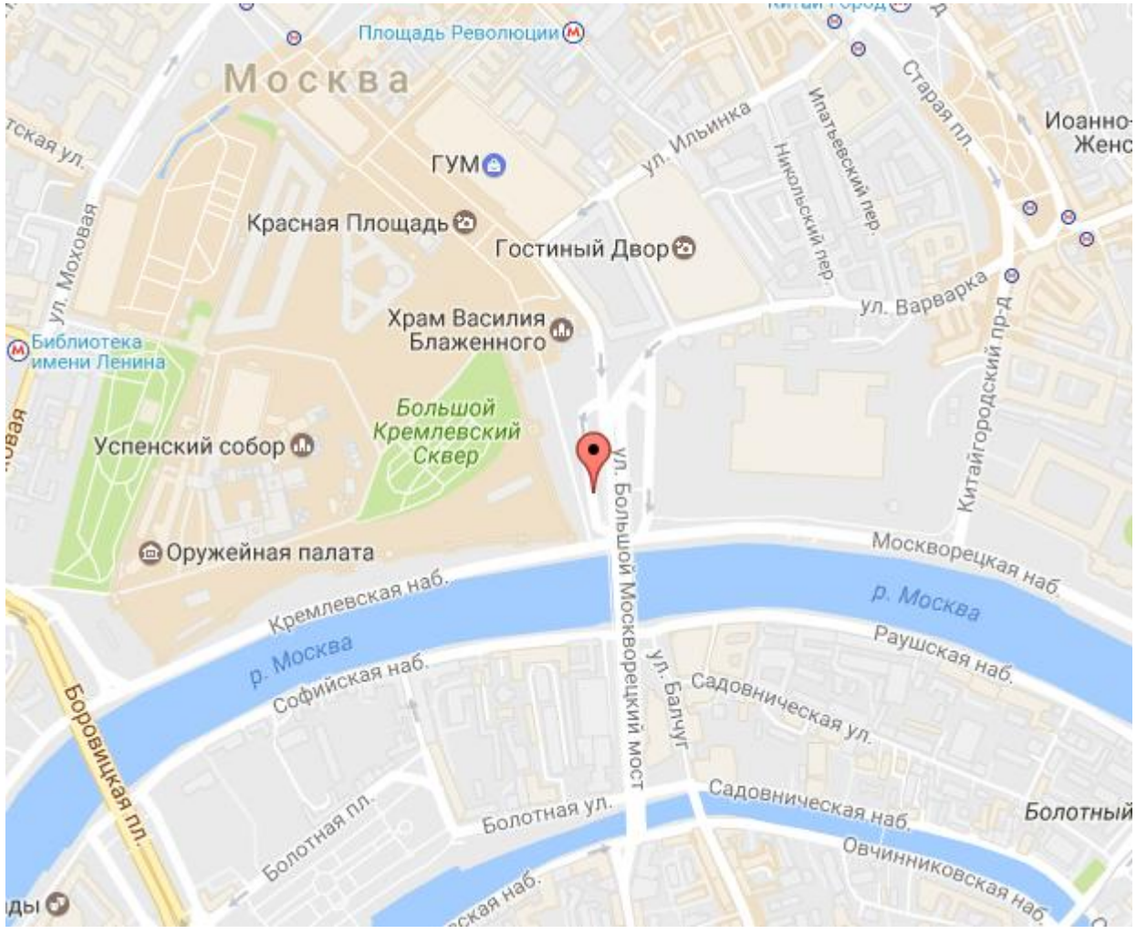
	merchant_id	latitude	longitude	transaction_time	record_time
5824	28477	0.000000	0.000000	2017-01-15 13:02:27	2017-01-15 13:02:20
5825	28477	0.000000	0.000000	2017-01-15 15:44:29	2017-01-15 15:54:15
5826	28477	0.000000	0.000000	2017-01-15 21:33:27	2017-01-15 21:38:17
5827	28477	0.000000	0.000000	2017-01-15 21:33:27	2017-01-15 21:39:21
5828	28477	55.211551	35.773620	2017-01-15 12:02:51	2017-01-15 11:59:56
5829	28477	52.593124	39.561907	2017-01-15 15:48:41	2017-01-15 15:49:49
5830	28477	51.178900	-1.826400	2017-01-15 17:05:51	2017-01-15 17:01:15
5831	28477	55.697067	37.553810	2017-01-15 16:14:25	2017-01-15 16:19:34
5832	28477	51.716180	39.175545	2017-01-15 17:08:23	2017-01-15 17:10:35
5833	28477	55.612360	37.607125	2017-01-15 14:00:34	2017-01-15 14:00:17
5834	28477	51.717860	39.177682	2017-01-15 16:00:21	2017-01-15 16:07:10
5835	28477	55.750347	37.623851	2017-01-15 18:11:40	2017-01-15 18:03:50
5836	28477	51.712188	39.174119	2017-01-15 18:34:36	2017-01-15 18:40:54
5837	28477	55.697067	37.553810	2017-01-15 22:14:20	2017-01-15 22:16:25
5838	28477	51.717669	39.178541	2017-01-15 20:30:28	2017-01-15 20:28:13
5839	28477	51.717268	39.177014	2017-01-15 22:57:16	2017-01-15 22:52:35
5840	28477	51.717867	39.177927	2017-01-15 19:34:17	2017-01-15 19:41:22
5841	28477	0.000000	0.000000	2017-01-15 15:44:29	2017-01-15 15:52:38
5842	28477	51.655555	39.153889	2017-01-15 10:57:44	2017-01-15 10:51:54
5843	28477	0.000000	0.000000	2017-01-15 18:02:27	2017-01-15 18:02:06

Визуализация данных – по какому принципу упорядочены данные?

	merchant_id	latitude	longitude	transaction_time	record_time
5824	28477	0.000000	0.000000	2017-01-15 13:02:27	2017-01-15 13:02:20
5825	28477	0.000000	0.000000	2017-01-15 15:44:29	2017-01-15 15:54:15
5826	28477	0.000000	0.000000	2017-01-15 21:33:27	2017-01-15 21:38:17
5827	28477	0.000000	0.000000	2017-01-15 21:33:27	2017-01-15 21:39:21
5828	28477	55.211551	35.773620	2017-01-15 12:02:51	2017-01-15 11:59:56
5829	28477	52.593124	39.561907	2017-01-15 15:48:41	2017-01-15 15:49:49
5830	28477	51.178900	-1.826400	2017-01-15 17:05:51	2017-01-15 17:01:15
5831	28477	55.697067	37.553810	2017-01-15 16:14:25	2017-01-15 16:19:34
5832	28477	51.716180	39.175545	2017-01-15 17:08:23	2017-01-15 17:10:35
5833	28477	55.612360	37.607125	2017-01-15 14:00:34	2017-01-15 14:00:17
5834	28477	51.717860	39.177682	2017-01-15 16:00:21	2017-01-15 16:07:10
5835	28477	55.750347	37.623851	2017-01-15 18:11:40	2017-01-15 18:03:50
5836	28477	51.712188	39.174119	2017-01-15 18:34:36	2017-01-15 18:40:54
5837	28477	55.697067	37.553810	2017-01-15 22:14:20	2017-01-15 22:16:25
5838	28477	51.717669	39.178541	2017-01-15 20:30:28	2017-01-15 20:28:13
5839	28477	51.717268	39.177014	2017-01-15 22:57:16	2017-01-15 22:52:35
5840	28477	51.717867	39.177927	2017-01-15 19:34:17	2017-01-15 19:41:22
5841	28477	0.000000	0.000000	2017-01-15 15:44:29	2017-01-15 15:52:38
5842	28477	51.655555	39.153889	2017-01-15 10:57:44	2017-01-15 10:51:54
5843	28477	0.000000	0.000000	2017-01-15 18:02:27	2017-01-15 18:02:06

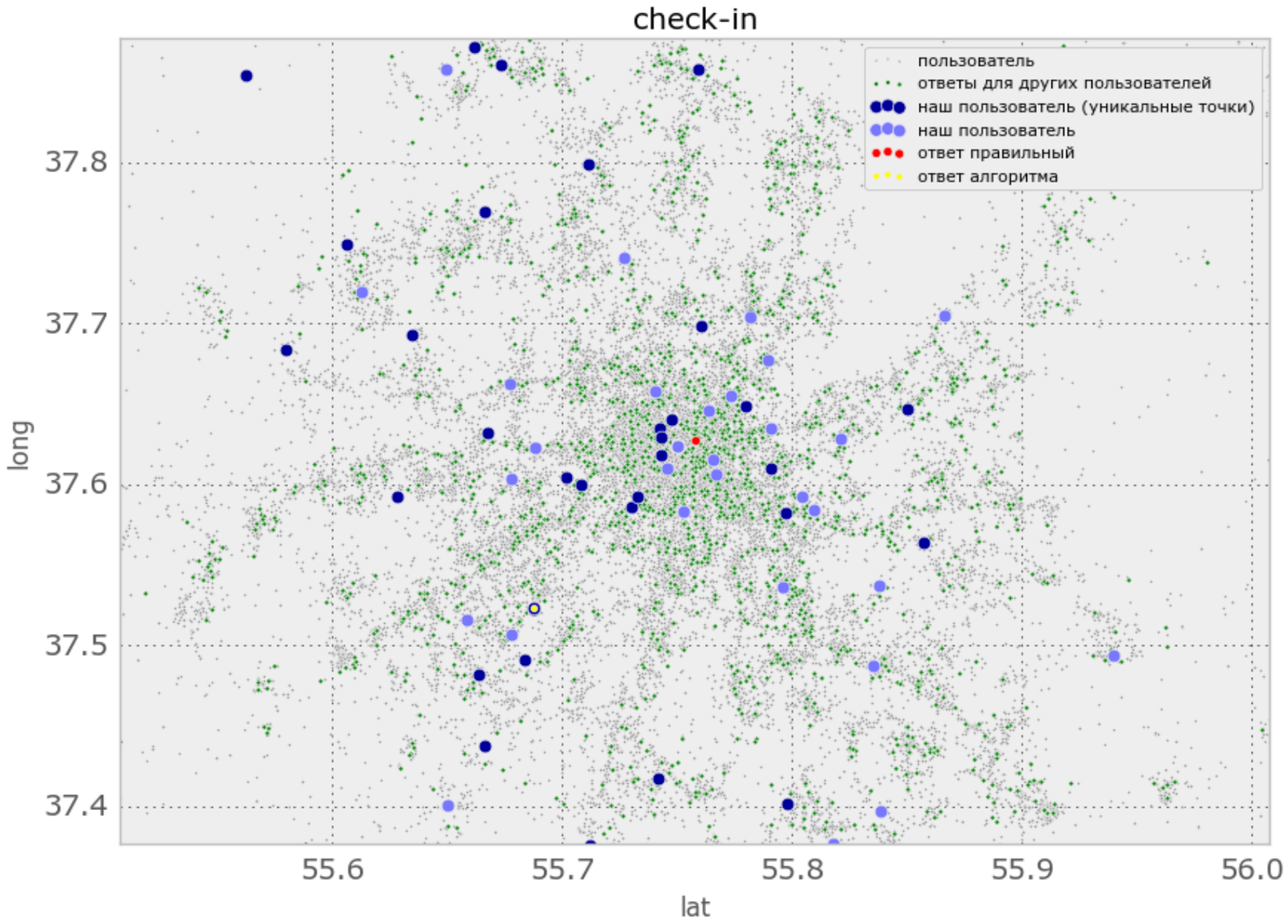
по дням… просто даты настоящих дней забиты «2017-01-15»

Визуализация данных – самый частый check-in

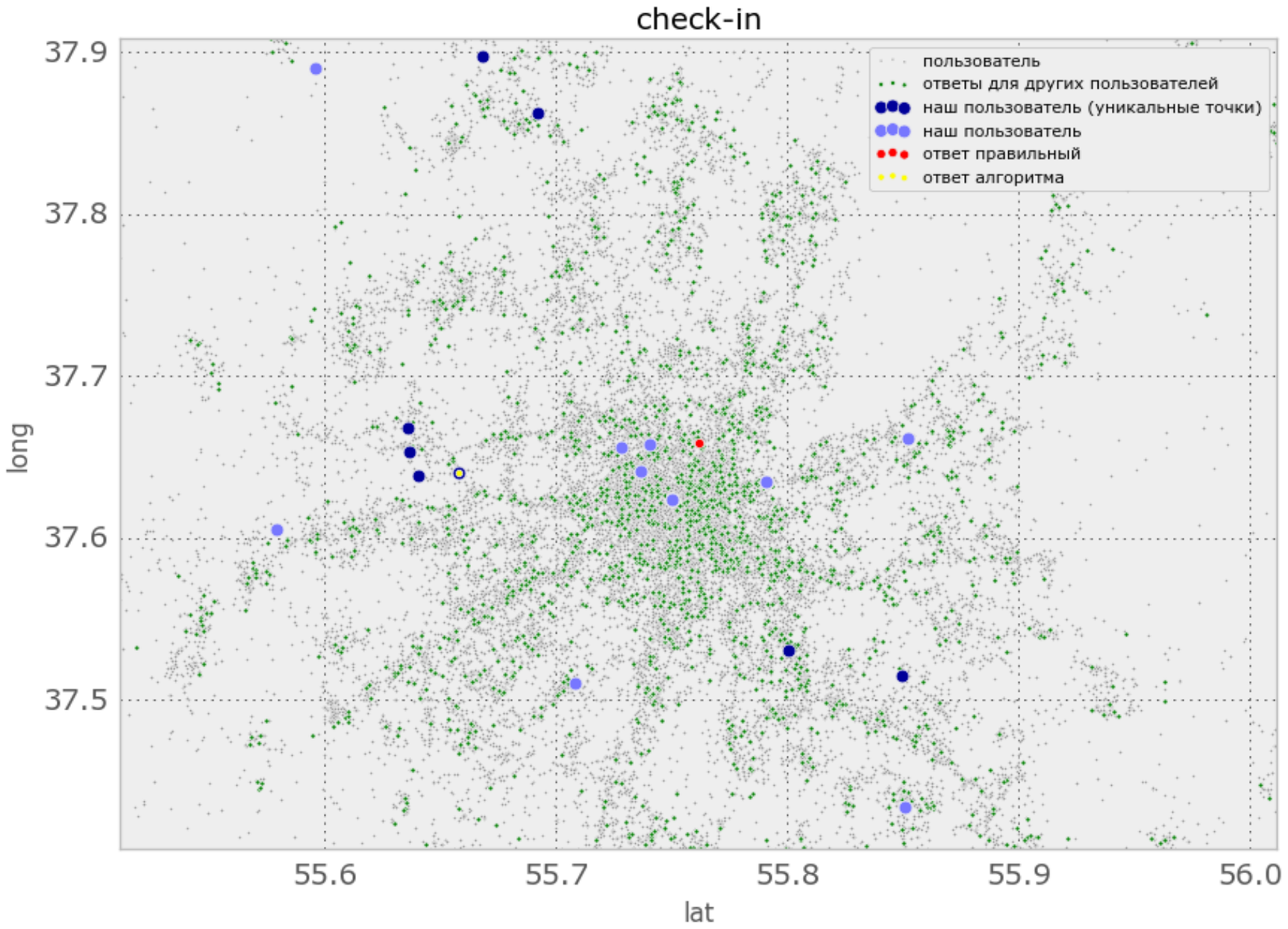


55.75034704 37.62385111 5321

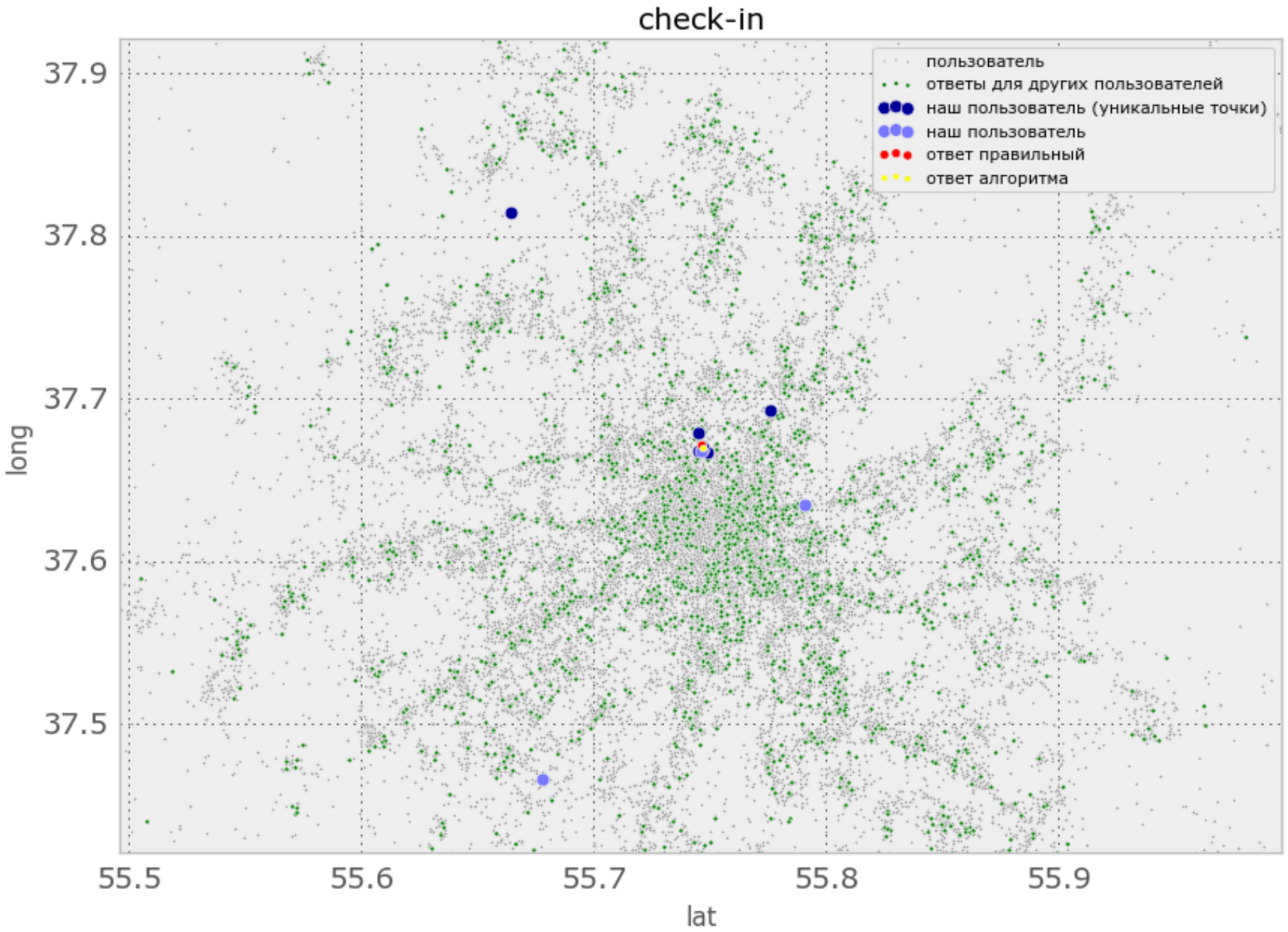
Визуализация данных – check-in



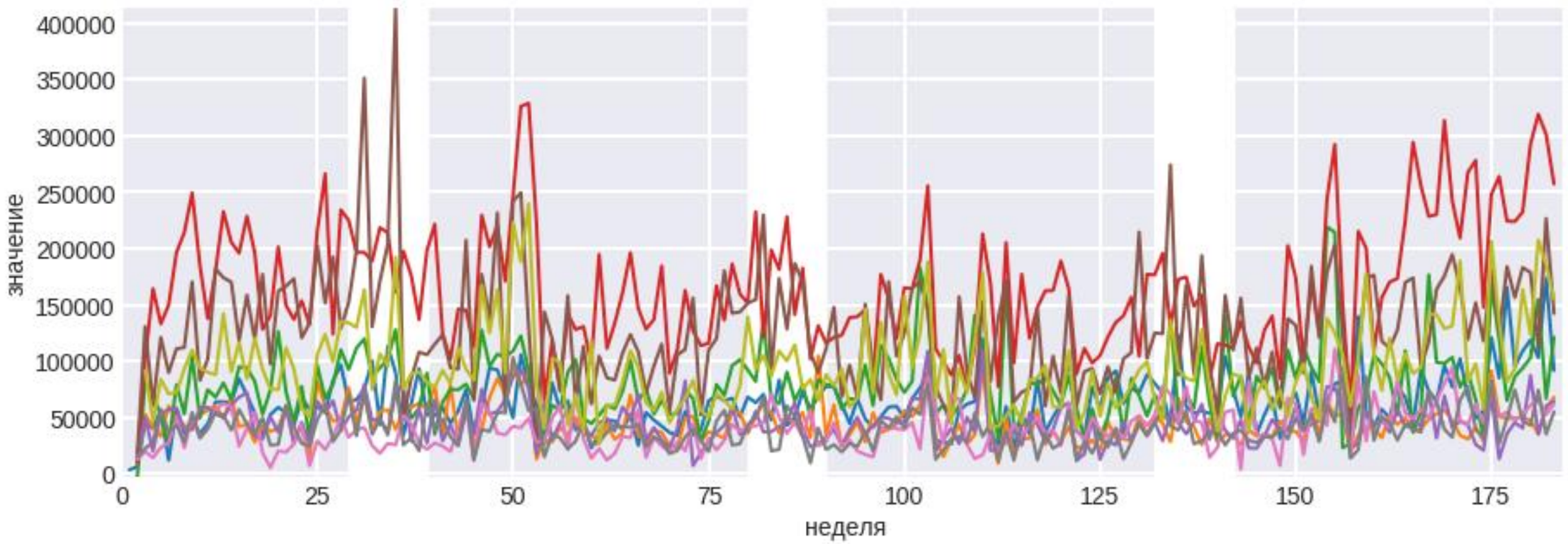
Визуализация данных – check-in



Визуализация данных – check-in



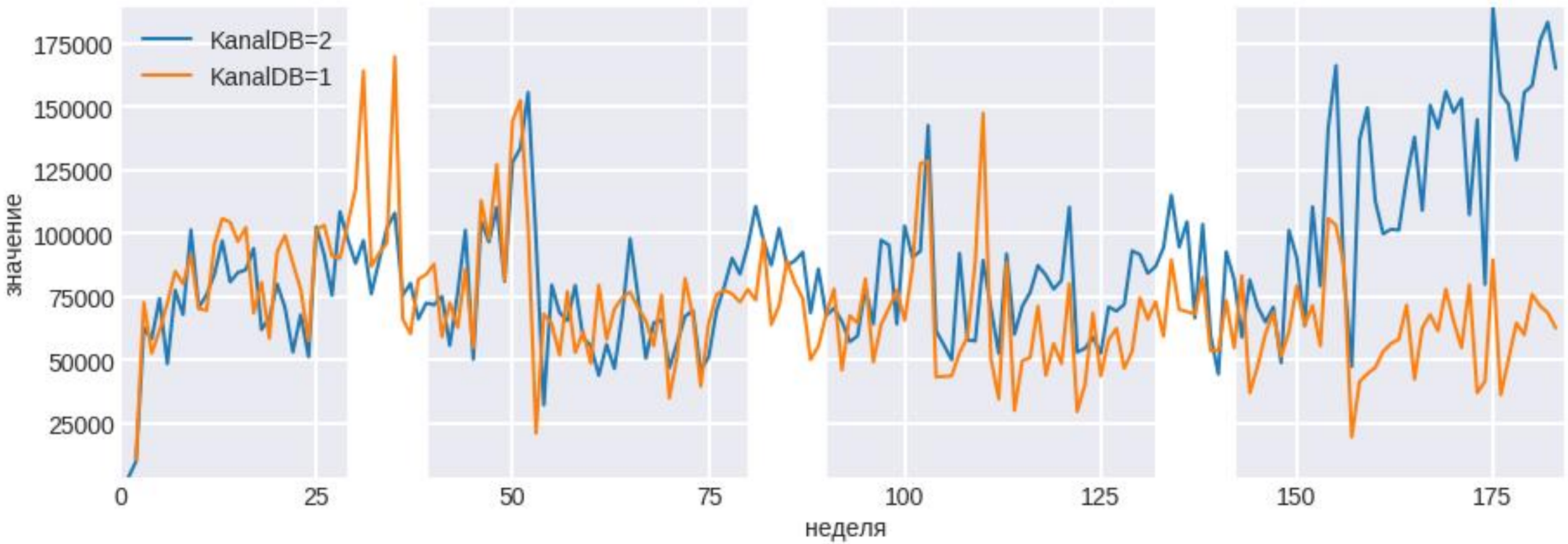
Визуализация данных – продажи Ascott Group



агрегаты продаж по разным каналам…

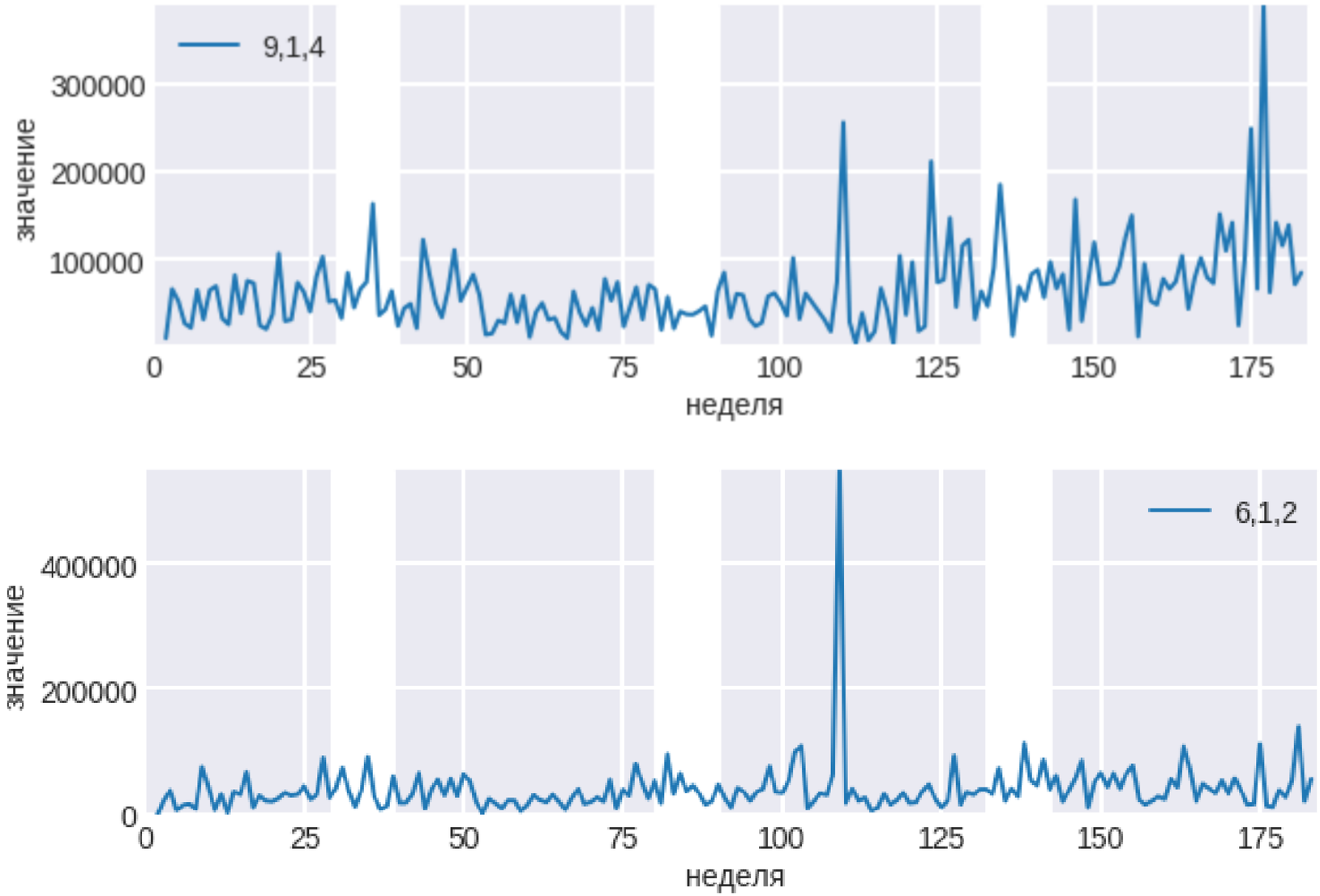
Белым показаны зоны, которые отстоят от зоны прогнозы на год, два и т.д.

Визуализация данных – продажи Ascott Group



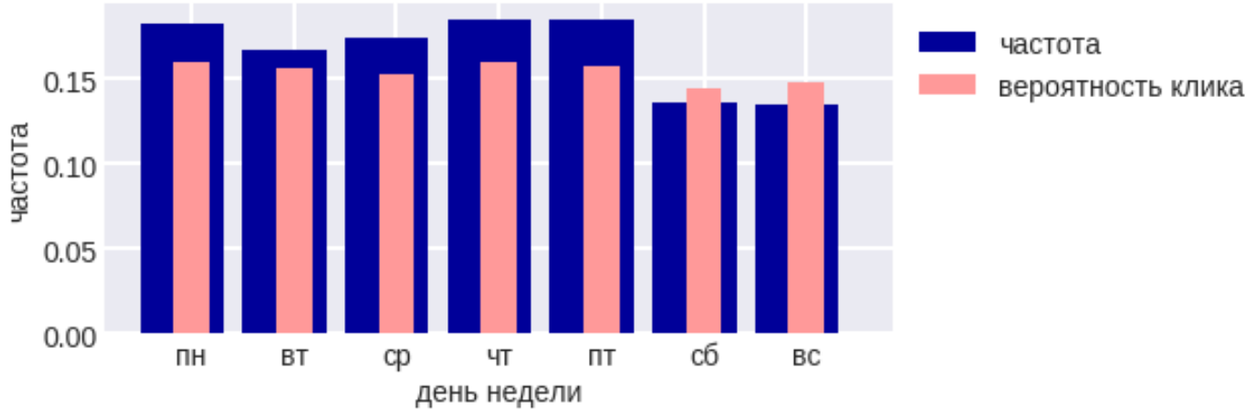
принципиально разные каналы!

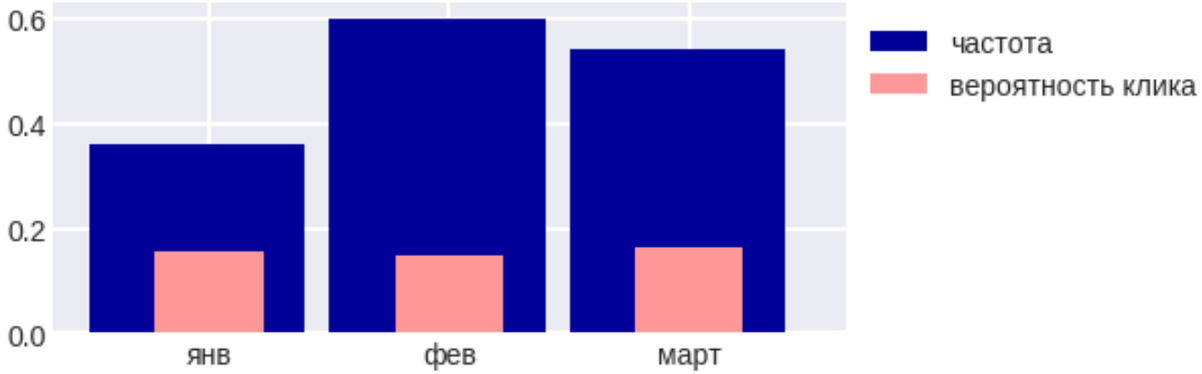
Визуализация данных – продажи Ascott Group

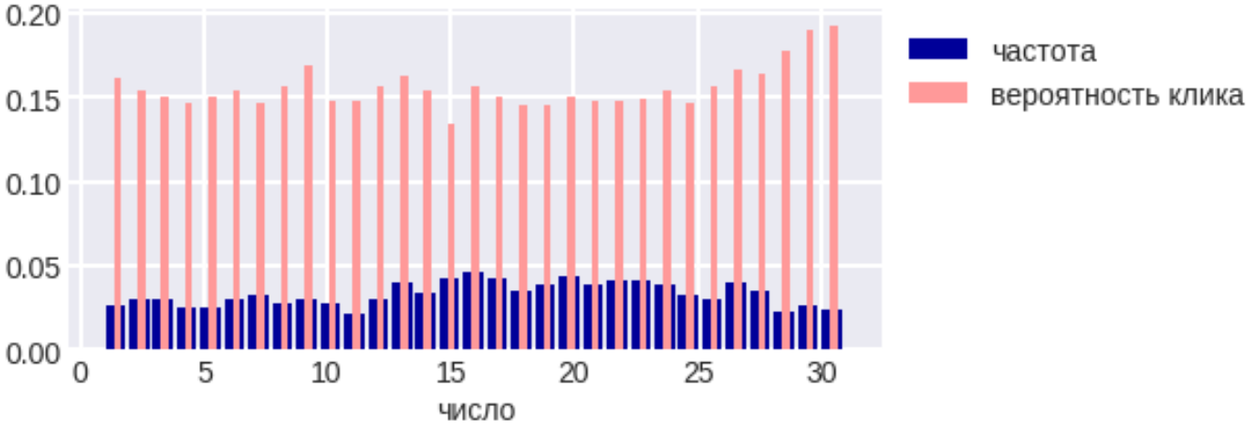


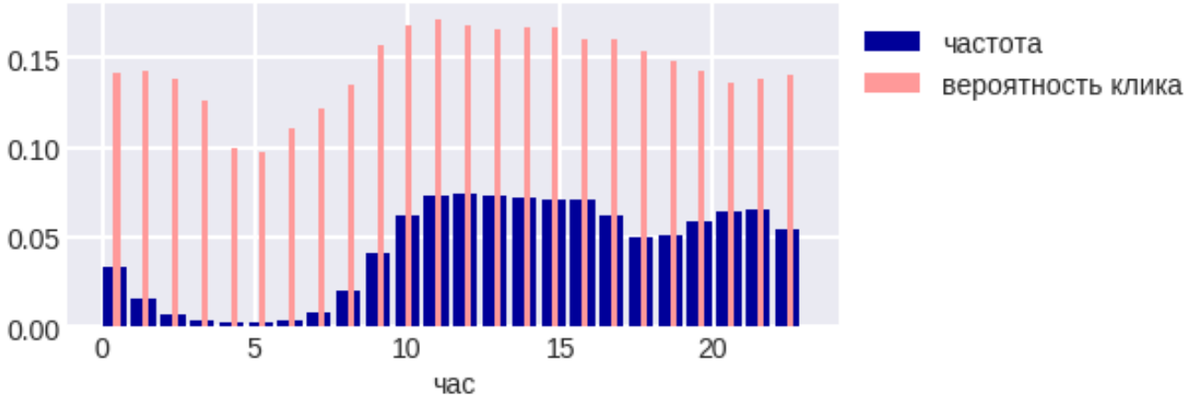
тренды и выбросы

Визуализация данных – Ticketland ML Contest



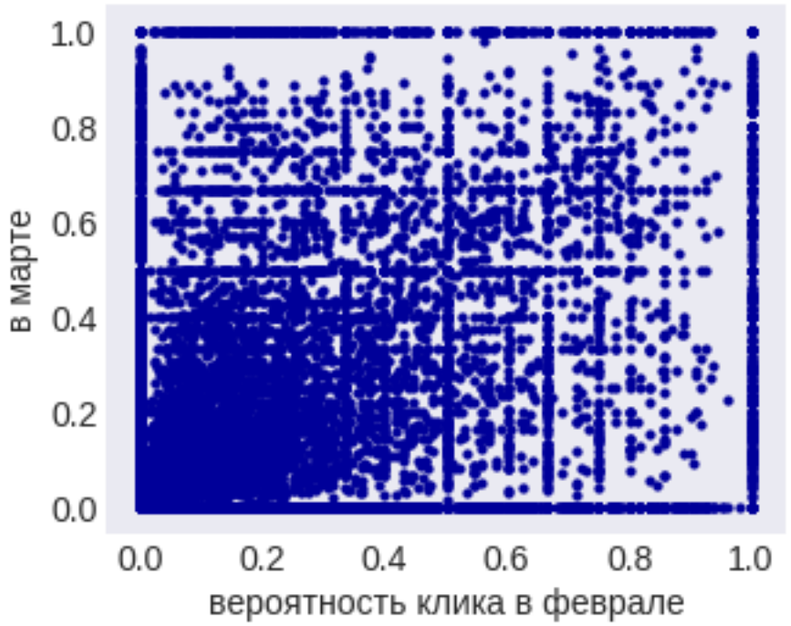


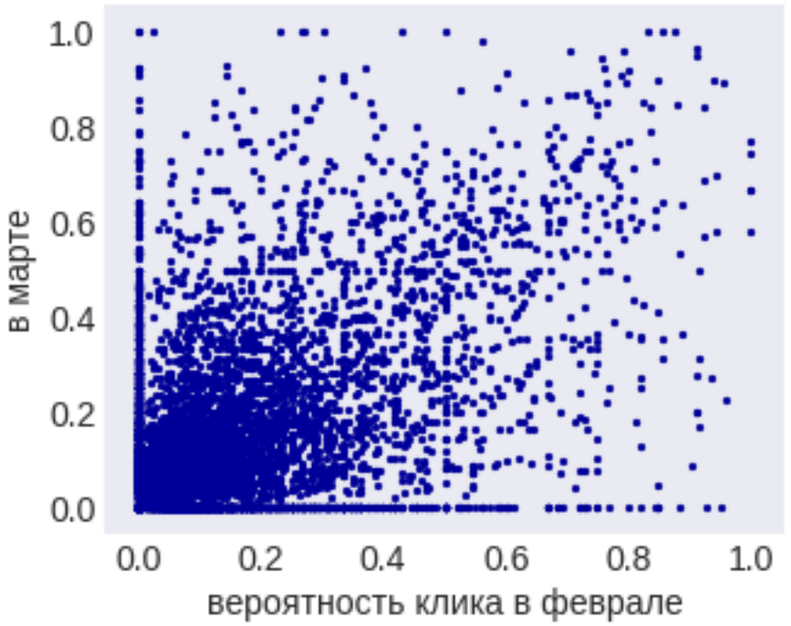




вероятность коррелирует с популярностью

Визуализация данных – Ticketland ML Contest



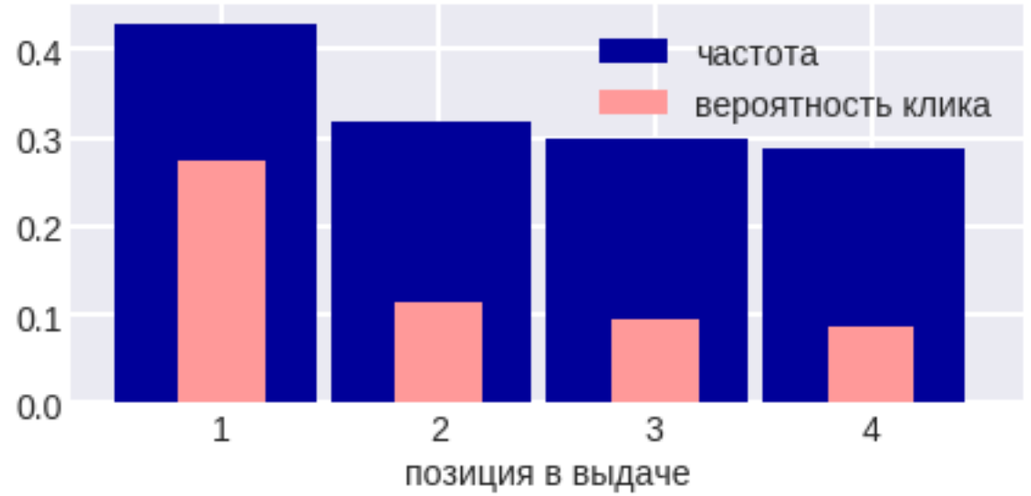


Справа – клиенты, которые наиболее активны

Визуализация данных – Ticketland ML Contest

Возраст: статистика по всем клиентам

есть артефакты (<0)



На первую выдачу чаще кликают!

Визуализация данных – Ticketland ML Contest

105000 показов, CTR 8%

82000 показов, CTR 6.5%

53000 показов, CTR 16%

16400 показов, CTR 91%

1900 показов, CTR 85%

8500 показов, CTR 83%

50000 показов, CTR 27.5%

47000 показов, CTR 6%

часто показываемые баннеры

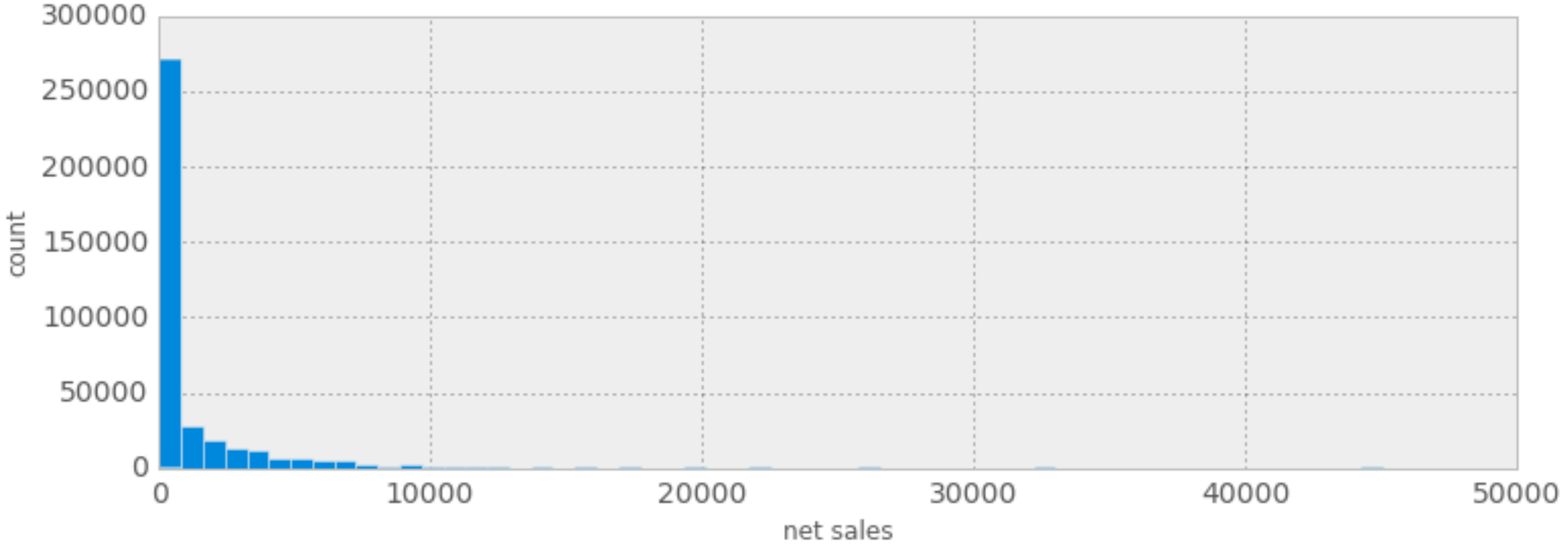
20000 показов, CTR 77%

1100 показов, CTR 76%

наиболее «кликабельные» баннеры

Визуализация данных – Ozon

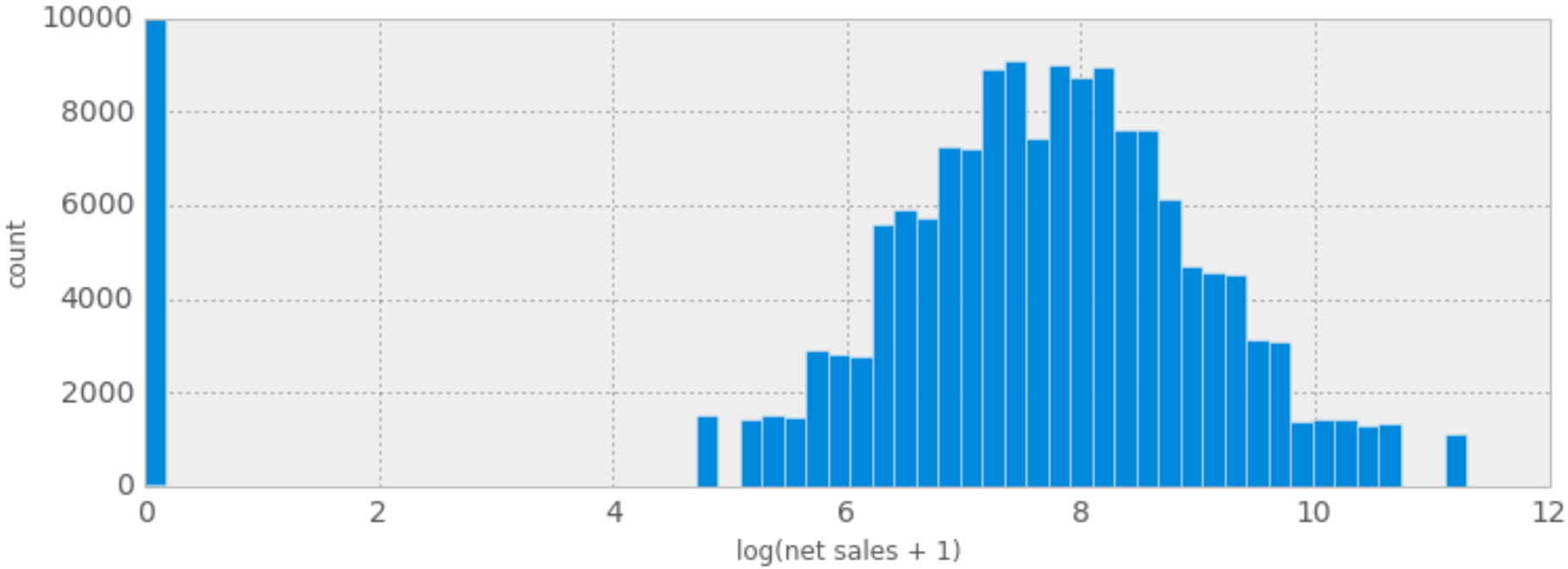
Распределение значений целевого признака



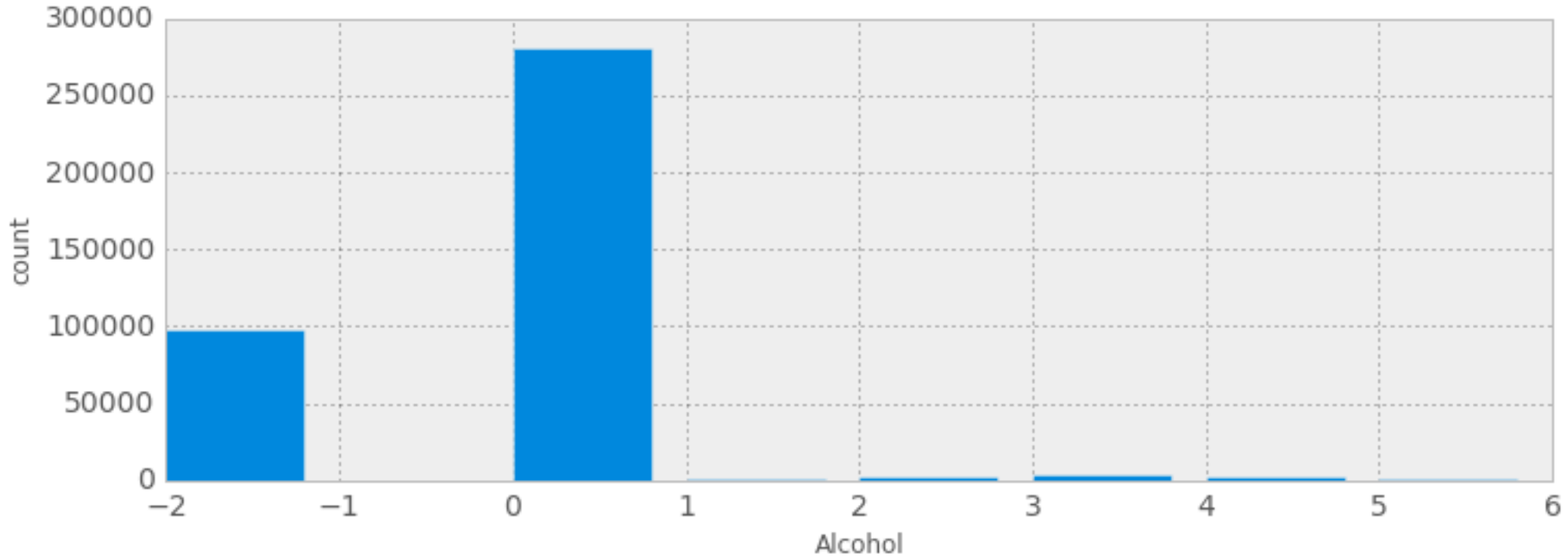
Предсказать ожидаемые продажи клиентов Озон за год после регистрации по их открытым данным в социальных сетях

Визуализация данных – Ozon

Распределение значений логарифма целевого признака



Визуализация данных – Ozon

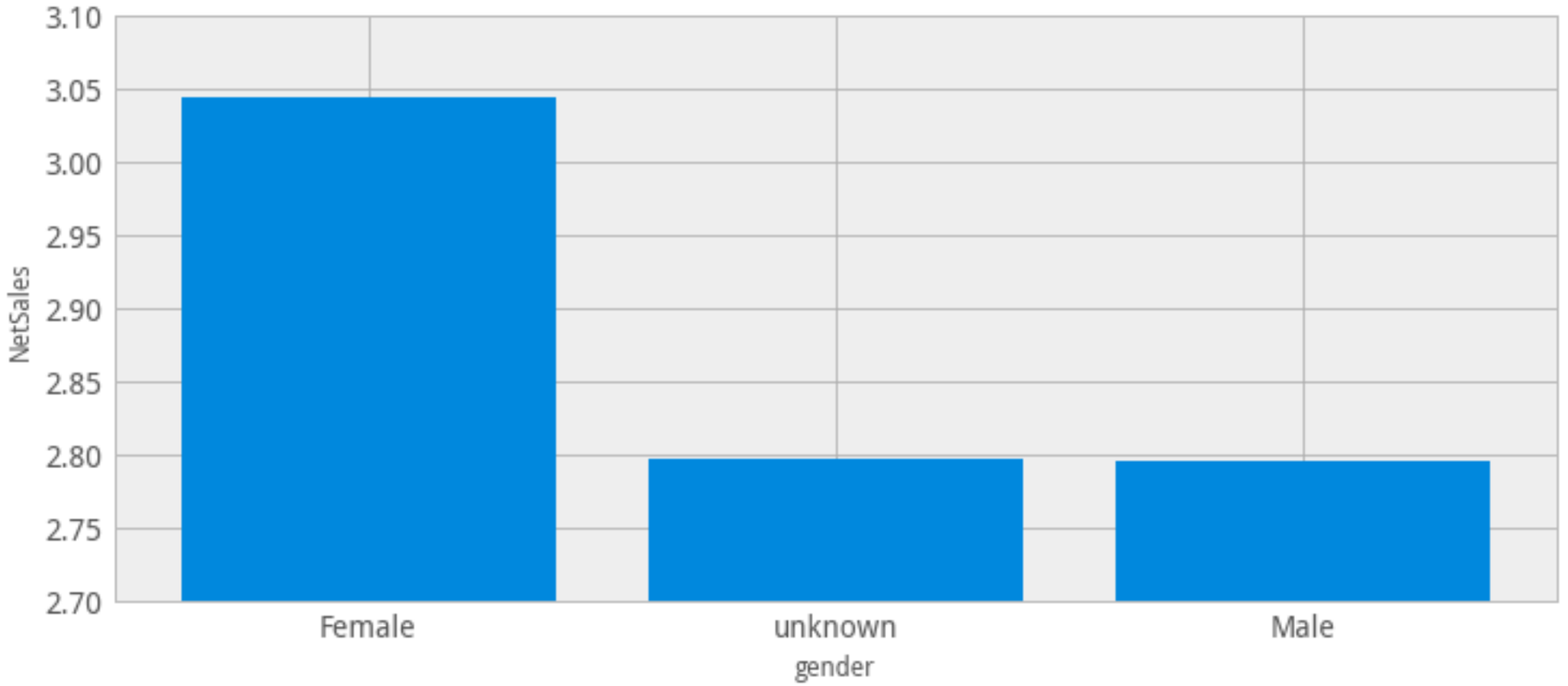


Отношение к алкоголю

Проблема – большинство значений неизвестно

и так почти у всех признаков

Визуализация данных – Ozon



Число значений признака по полу

Следовательно, пол «не указывают» в основном мужчины

Визуализация данных – Ozon

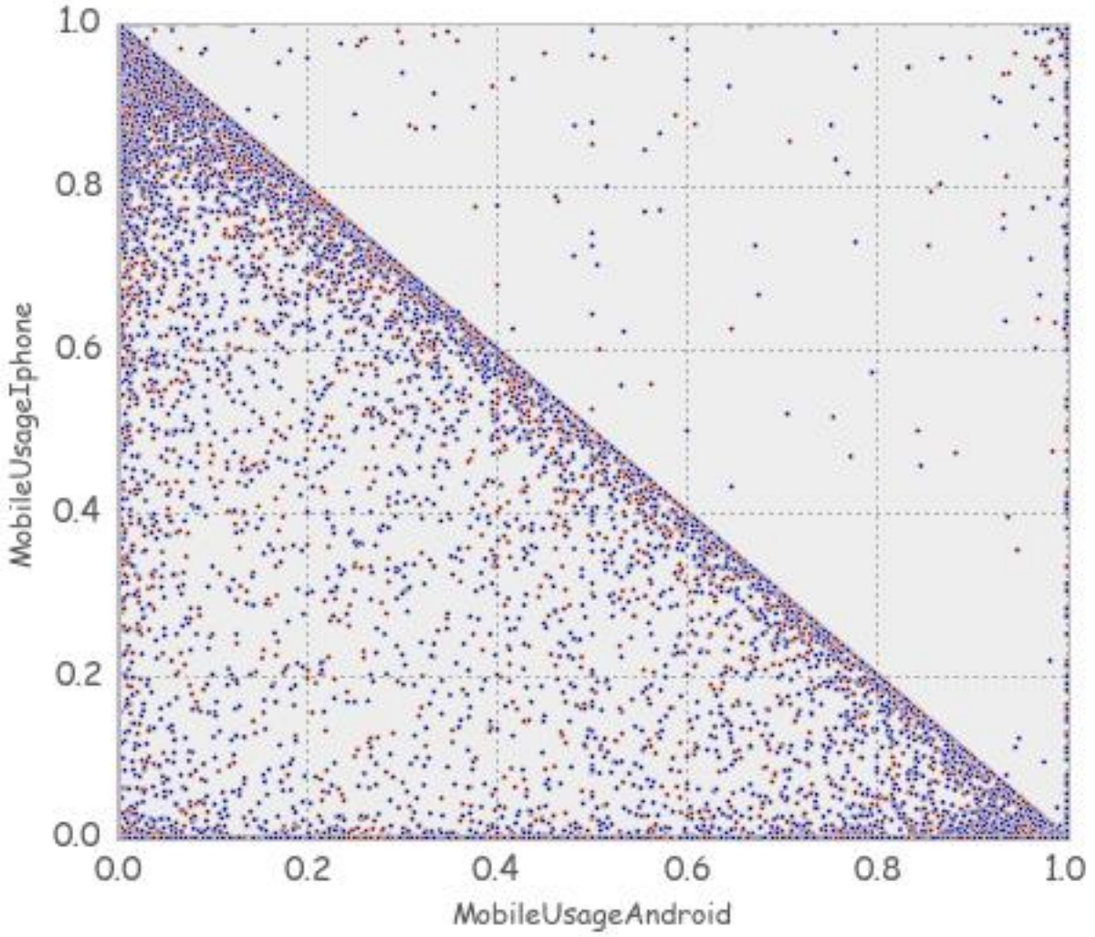
HasPhone	HasSkype	HasTwitter	count	mean
-2	-2	-2	97362	2.937212
0	0	0	253128	2.994881
0	0	1	1312	2.807146
0	1	0	16811	2.872637
0	1	1	1589	2.796710
1	0	0	11847	2.460538
1	0	1	246	2.337781
1	1	0	6959	2.276721
1	1	1	746	2.445111

Наличие телефона, скайпа и твиттера

чем больше человек указывает информации в соцсети,

тем он более плохой покупатель…

Визуализация данных – Ozon



Разным цветом – положительность NetSales

MobileUsageAndroid – доля входов в аккаунт с устройств Android

MobileUsageIphone – доля входов в аккаунт с устройств IPhone

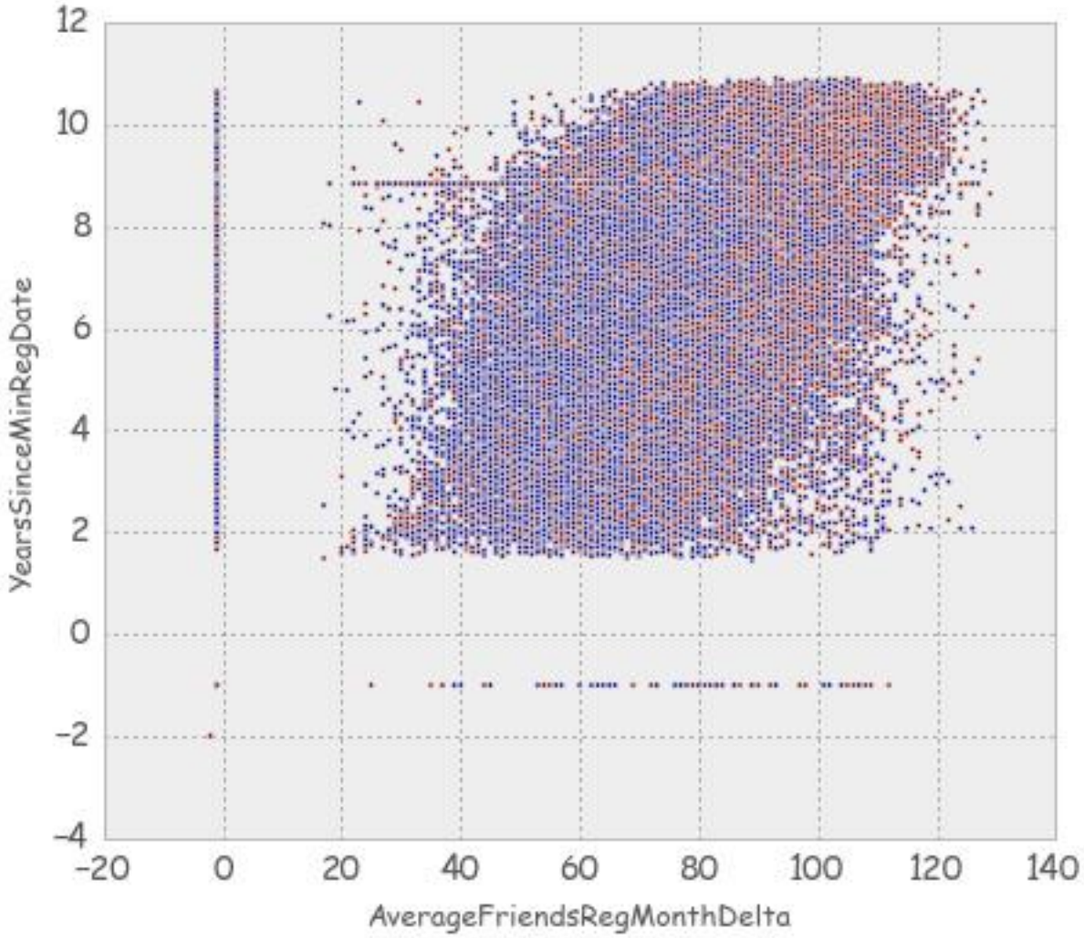
Визуализация данных – Ozon

Понятно, что сумма долей не может быть > 1

Интересно, что есть пользователи, которые пользуются Android и IPhone (в самых разных пропорциях!)

Опять же… для 64.8% пользователей ничего не известно…

Визуализация данных – Ozon



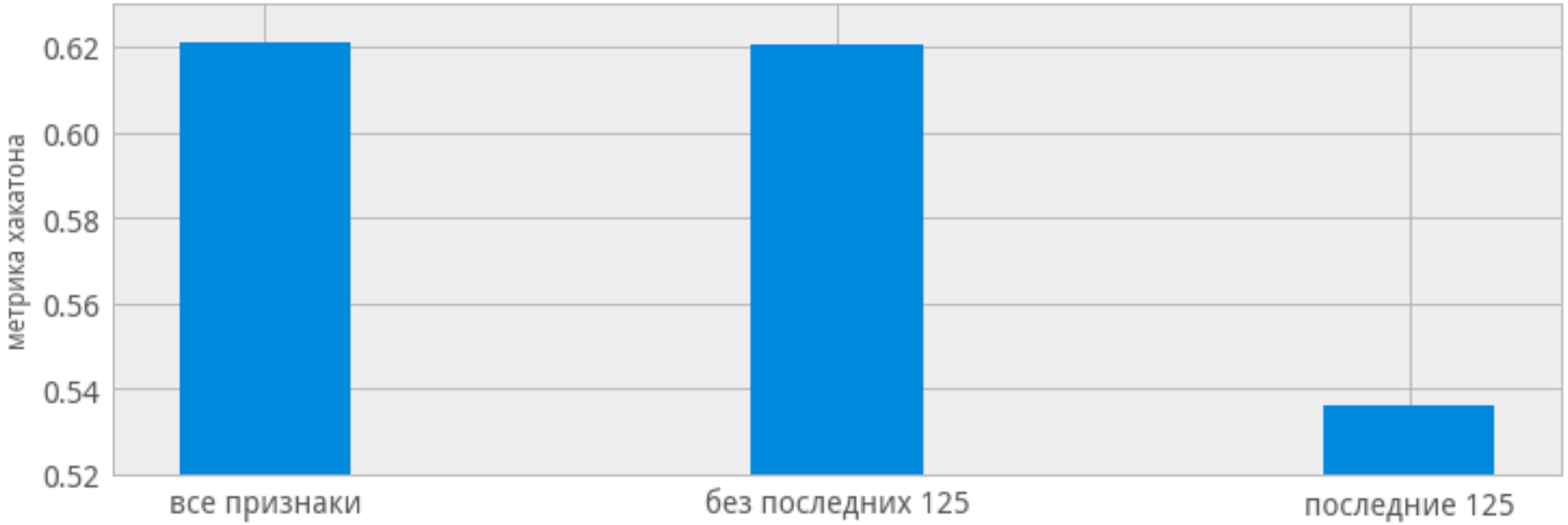
Два самых важных признака… AUC ~ 0.6

Визуализация данных – Ozon

AverageFriendsRegMonthDelta – средняя разница между текущей датой и датой регистрации всех друзей человека в соцсетях

YearsSinceMinRegDate - количество лет, прошедших с даты регистрации первого аккаунта в социальной сети

Не было признака «возраст», возможно, он самый важный!



Визуализация данных – Ozon

качество упирается в порог 62%

есть волшебные признаки (возрастные)

признаки интересов бесполезны

очень много неизвестных признаков

(видимо, из-за некачественного парсинга соцсетей)

значения некоторых признаков некорректны

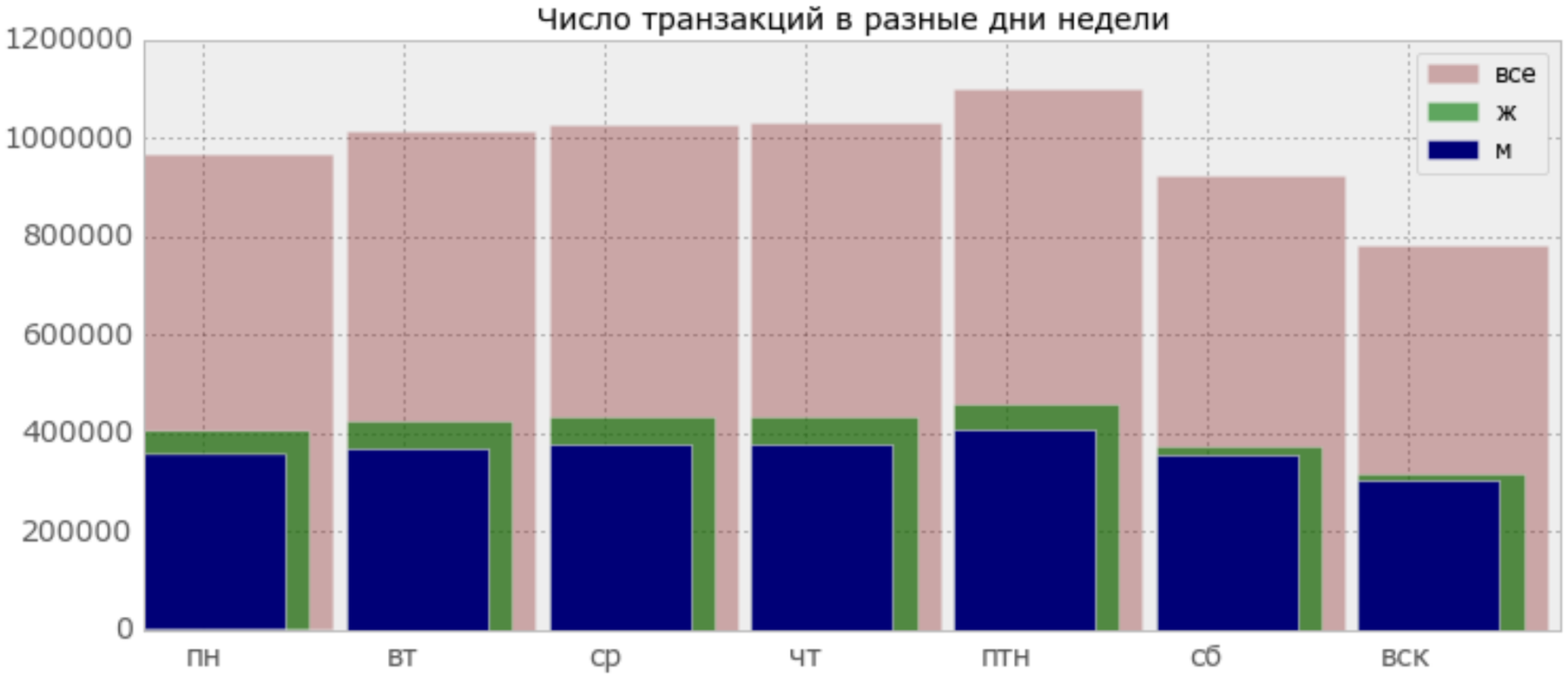
если бы все значения признаков были известны, качество превышало бы 65%

для решения лучше использовать Light GBM

- по данным соцсетей можно делать косвенные выводы

(для решения нашей задачи это бесполезно)

Визуализация данных – Сбербанк



все = м + ж + неизвестно

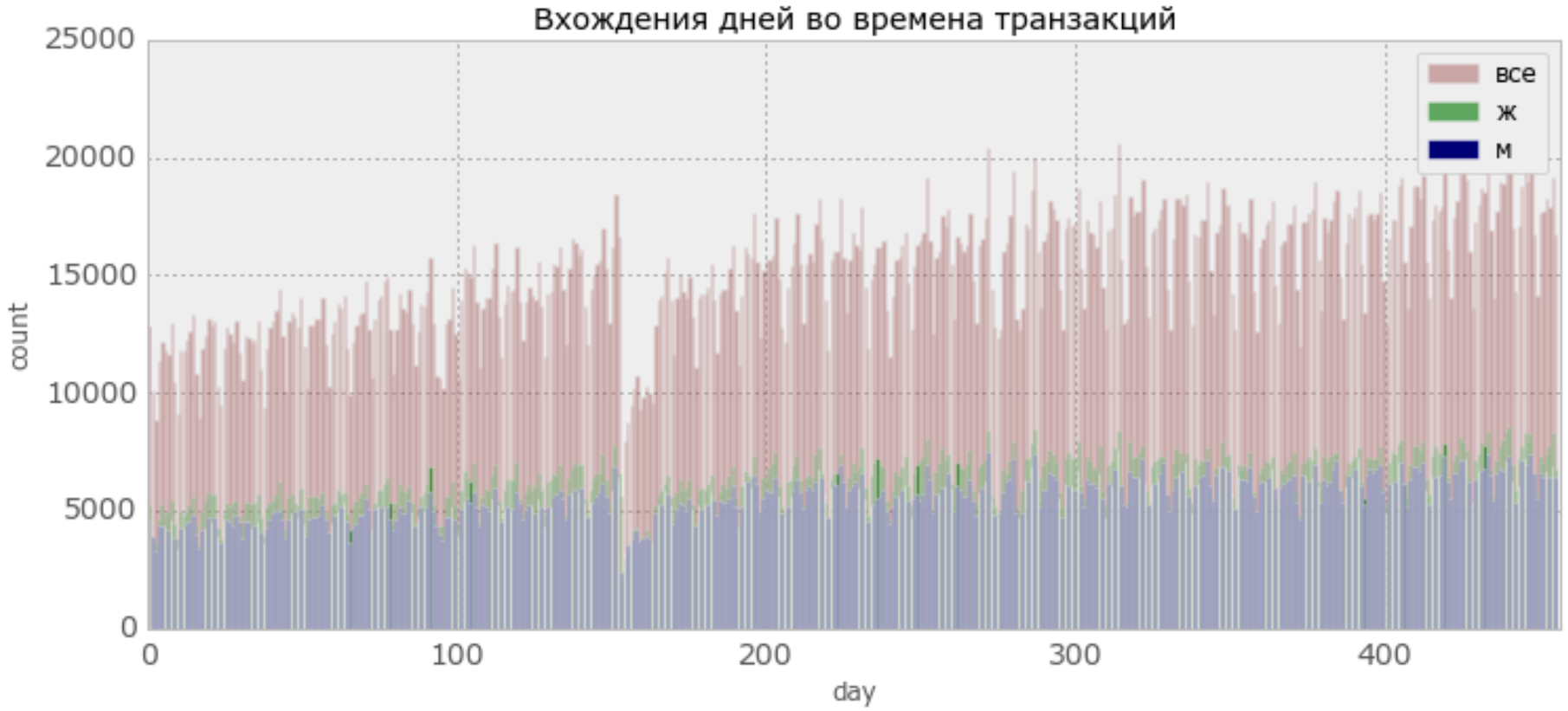
Задача: определить пол по истории транзакций

Визуализация данных – Сбербанк

В целом соответствует «естественному трудовому дню»…

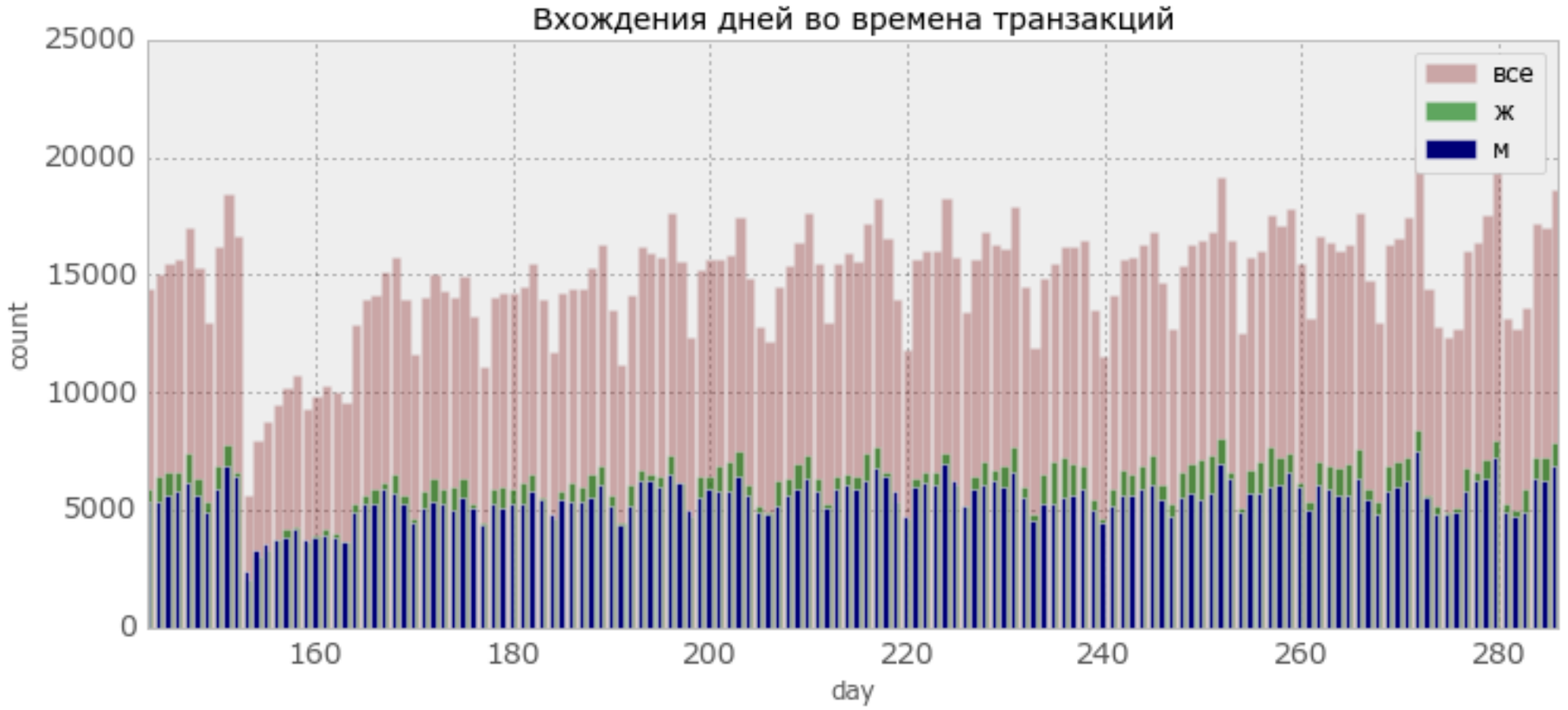
Было кодовое время «00:00:00»

Визуализация данных – Сбербанк



Провал 11 дней идентифицирует начало года

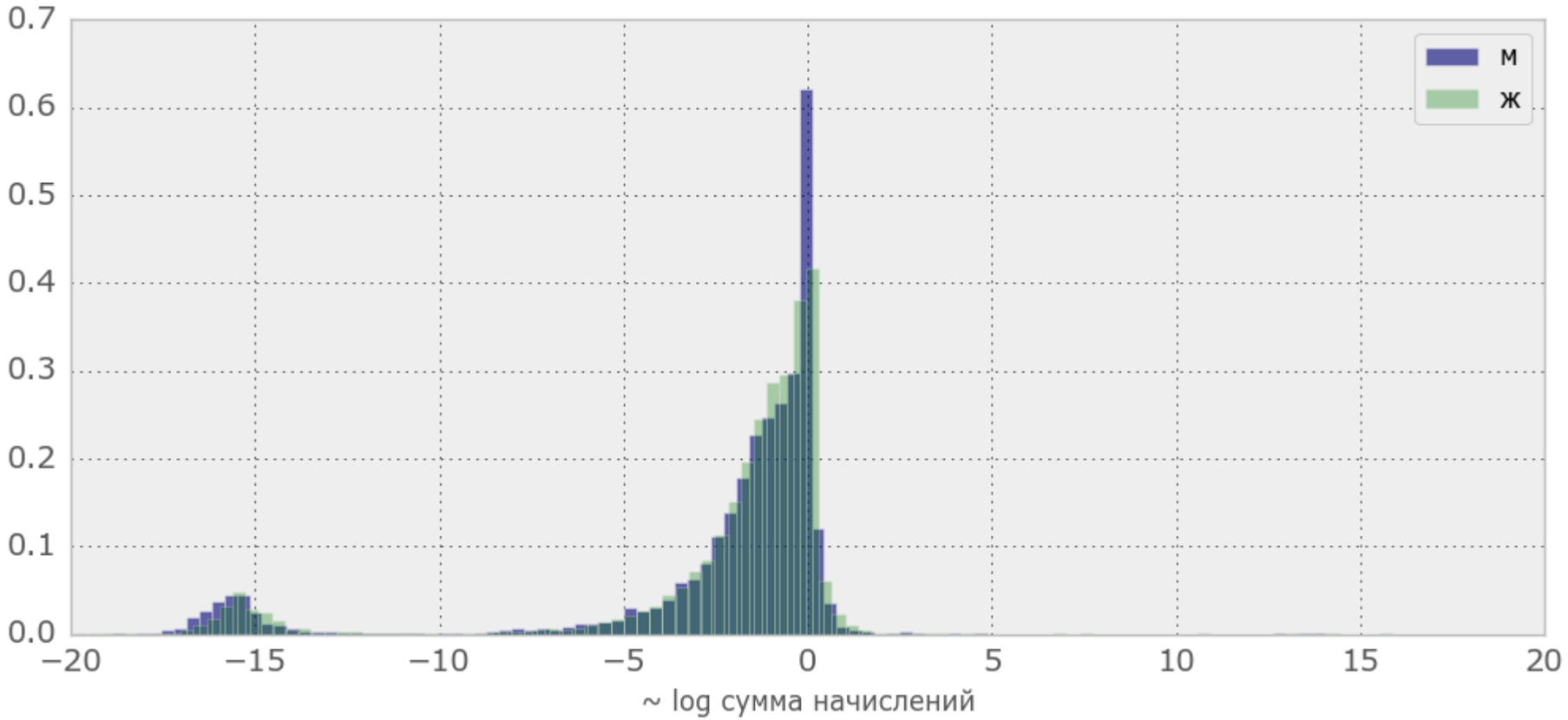
Визуализация данных – Сбербанк



Есть провалы на майские праздники

7-дневная цикличность

Визуализация данных – Сбербанк



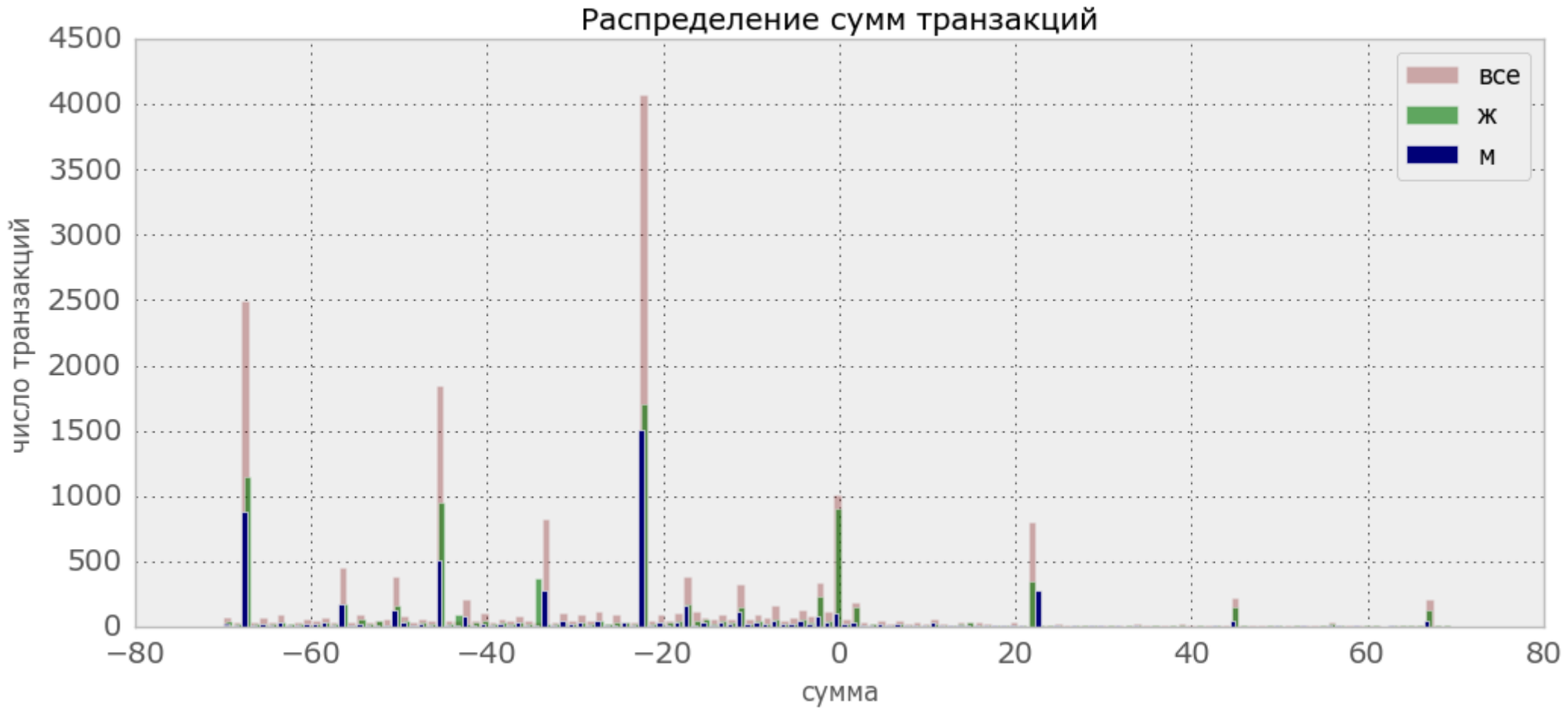
Логарифм разницы между начислениями и снятиями

Мужчины снимают всё!

Визуализация данных – Сбербанк

mcc_code	ж	м	mcc_description	k
5967	5	289	Прямой маркетинг — входящий телемаркетинг	0.97
5931	335	39	Магазины second hand, магазины б/у товаров, ко...	0.80
1731	8	65	Подрядчики по электричеству	0.78
7995	2431	15650	Транзакции по азартным играм	0.73
7994	1164	7404	Галереи/учреждения видеоигр	0.728
9211	43	7	Судовые выплаты, включая алименты и детскую по...	0.72
6211	133	776	Ценные бумаги: брокеры/дилеры	0.71
7512	22	123	Прокат автомобилей	0.697
5965	106	19	Прямой маркетинг — комбинированный каталог и т...	0.696
7993	106	591	Принадлежности для видеоигр	0.6958

Визуализация данных – Сбербанк



суммы были изменены

Неужели есть «Женские суммы трат»?!

Итог

увидели полезность приёмов визуализации,
которые обсуждали в основных лекциях