
Прикладные задачи анализа данных
Визуализация данных соревнования Сбербанка «Data Science Contest»
Александр Дьяконов
Московский государственный университет имени М.В. Ломоносова

Данные

Транзакции

	customer_id	tr_datetime	mcc_code	tr_type	amount	term_id
0	39026145	0 10:23:26	4814	1030	-2245.92	NaN
1	39026145	1 10:19:29	6011	7010	56147.89	NaN
2	39026145	1 10:20:56	4829	2330	-56147.89	NaN
3	39026145	1 10:39:54	5499	1010	-1392.47	NaN
4	39026145	2 15:33:42	5499	1010	-920.83	NaN

~ 7 млн

Целевой признак

	customer_id	gender
0	75562265	0
1	10928546	1
2	69348468	1
3	84816985	1
4	61009479	0

~ 12 000

mcc-коды

	mcc_code	mcc_description
0	742	Ветеринарные услуги
1	1711	Генеральные подрядчики по вентиляции, теплосна...
2	1731	Подрядчики по электричеству
3	1799	Подрядчики, специализированная торговля — нигд...
4	2741	Разнообразные издательства/печатное дело

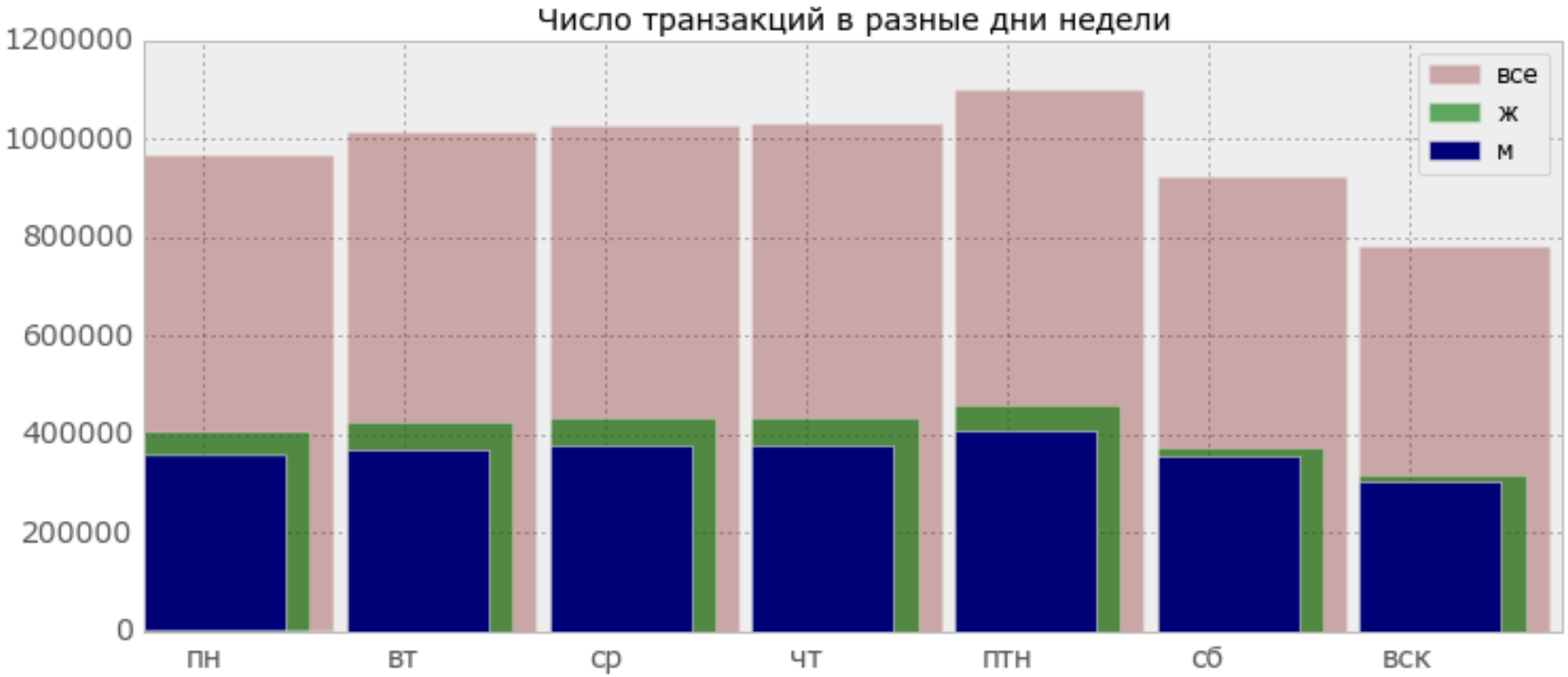
~ 184

Типы

	tr_type	tr_description
0	3200	Плата за предоставление услуг поср
1	3210	Плата за предоставление отчета по
2	3800	Плата за обслуживание банковской к
3	4000	Плата за получение наличных в Сбе
4	4001	Плата за получение наличных в Сбе

~ 155 (77)

Дни недели



Проставление дням меток сделана из эвристических соображений

все = м + ж + неизвестно

Секунды

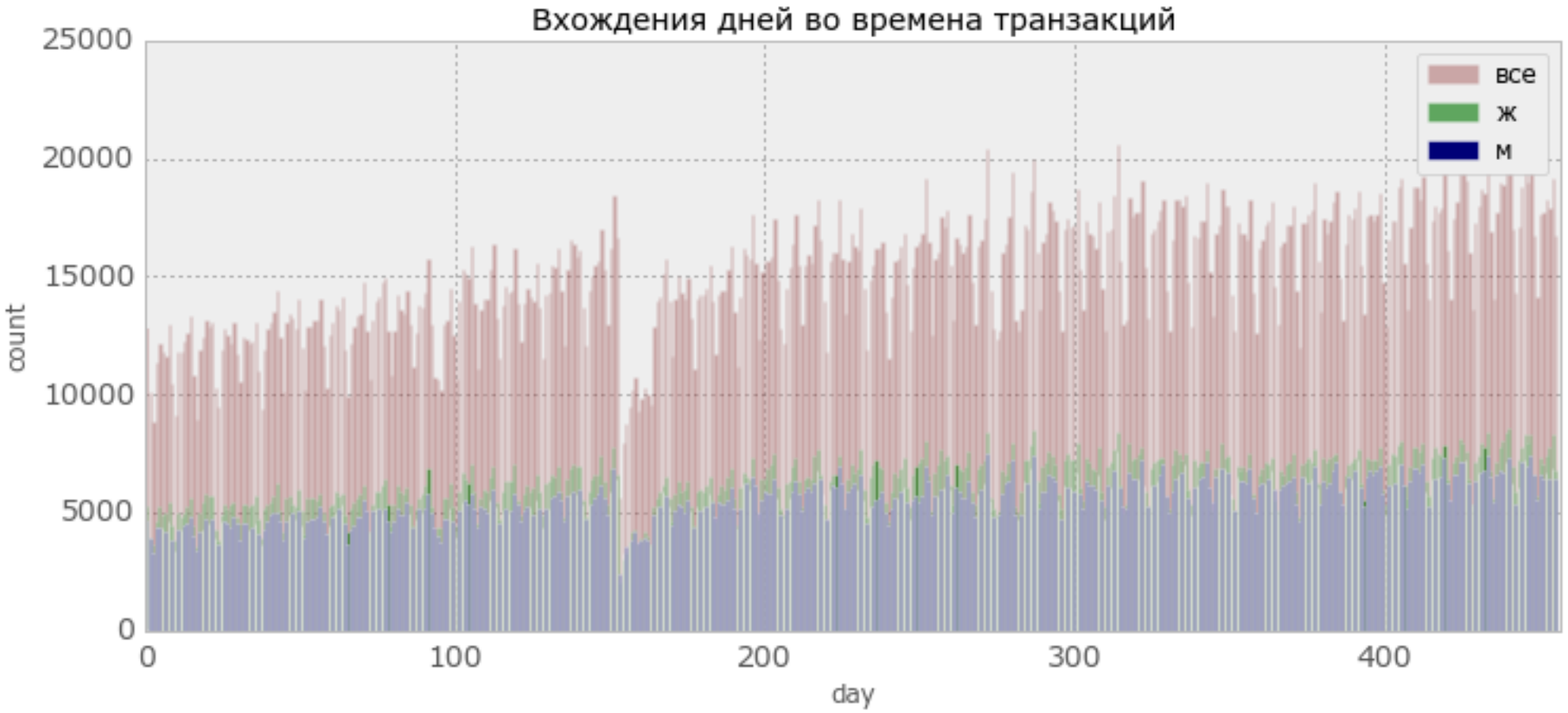
Особенность: 61 секунда (от 0 до 60)

Часто встречается время 00:00:00

Часы

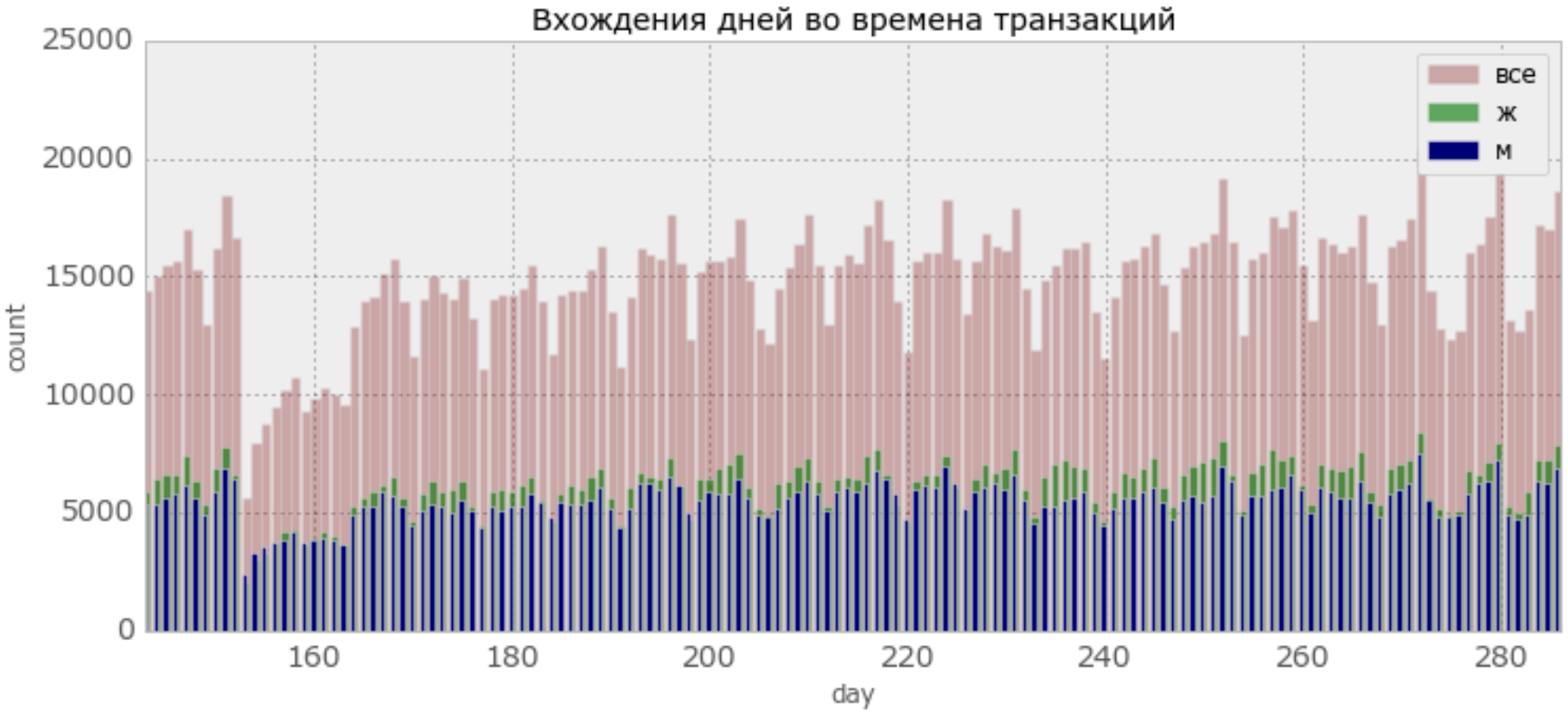
В целом соответствует «естественному трудовому дню»...

Дни



Провал 11 дней идентифицирует начало года

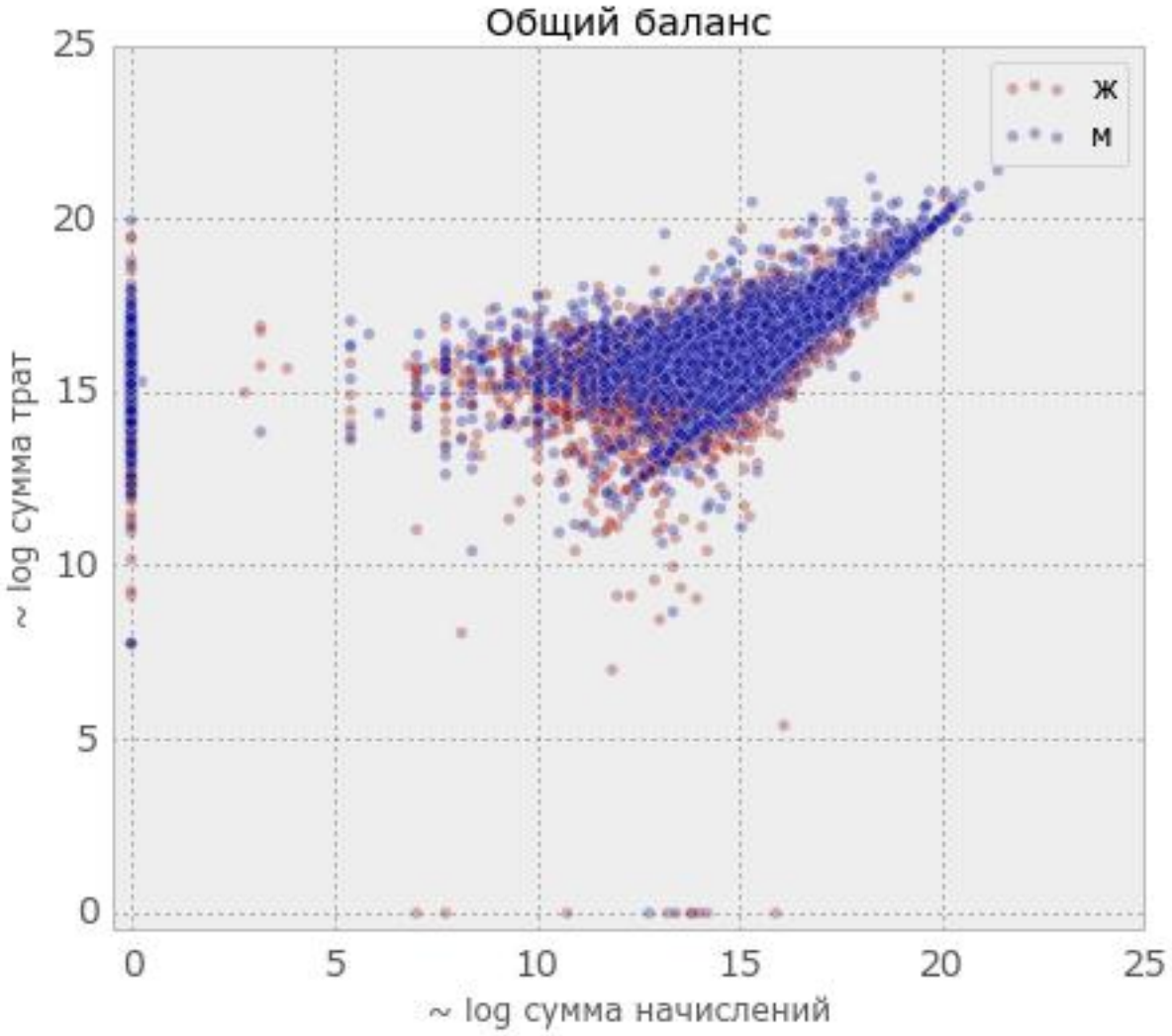
Дни



Есть провалы на майские праздники

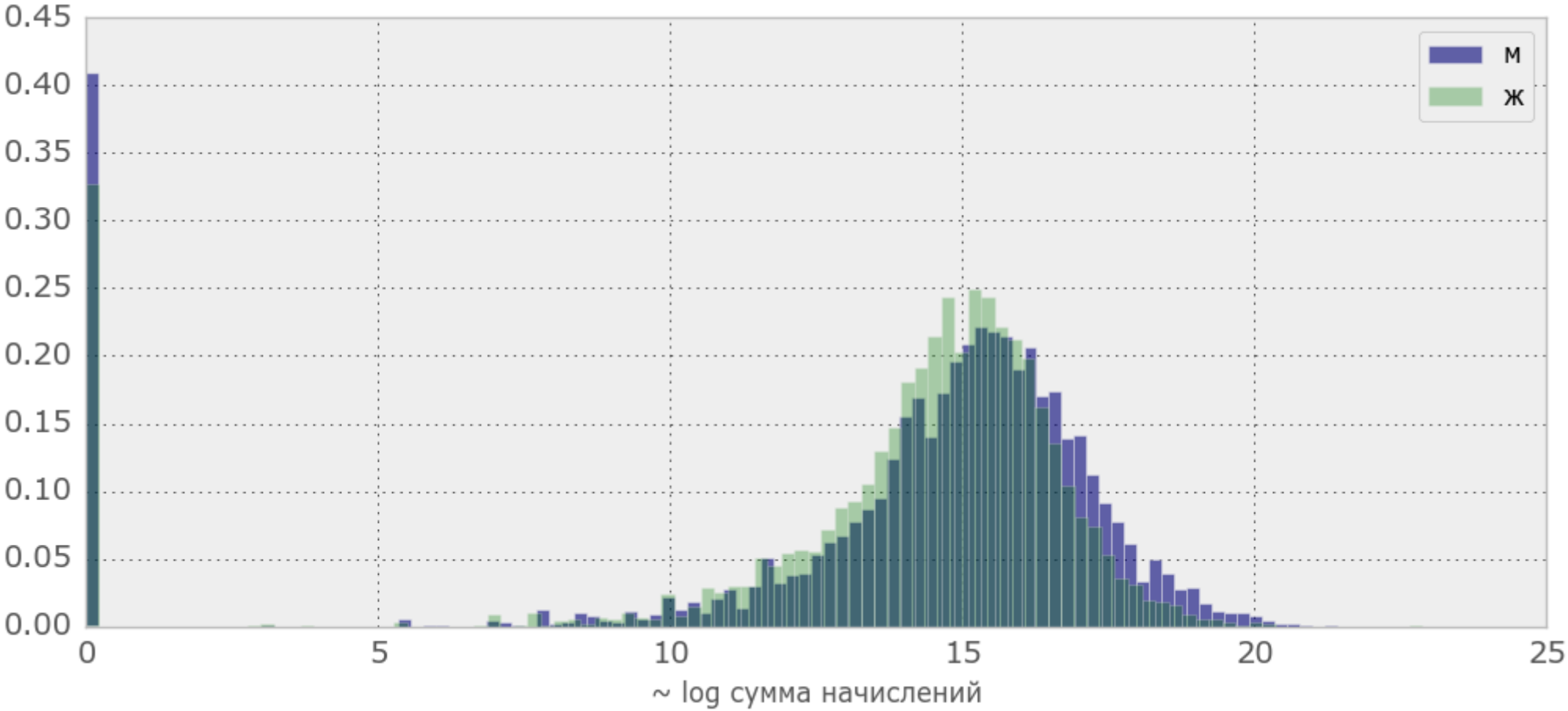
7-дневная цикличность

Суммы операций



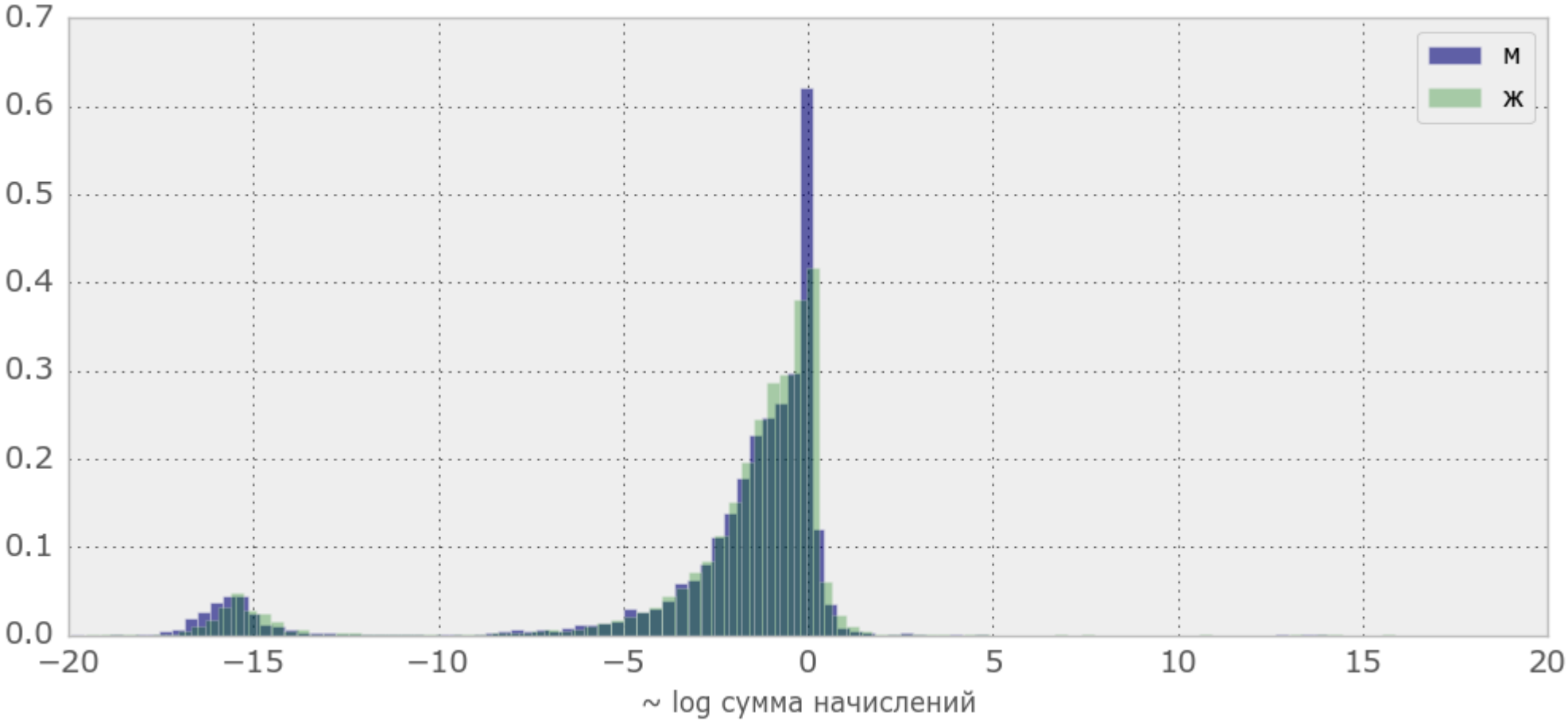
Видна линия равенства доходов и расходов

Суммы операций



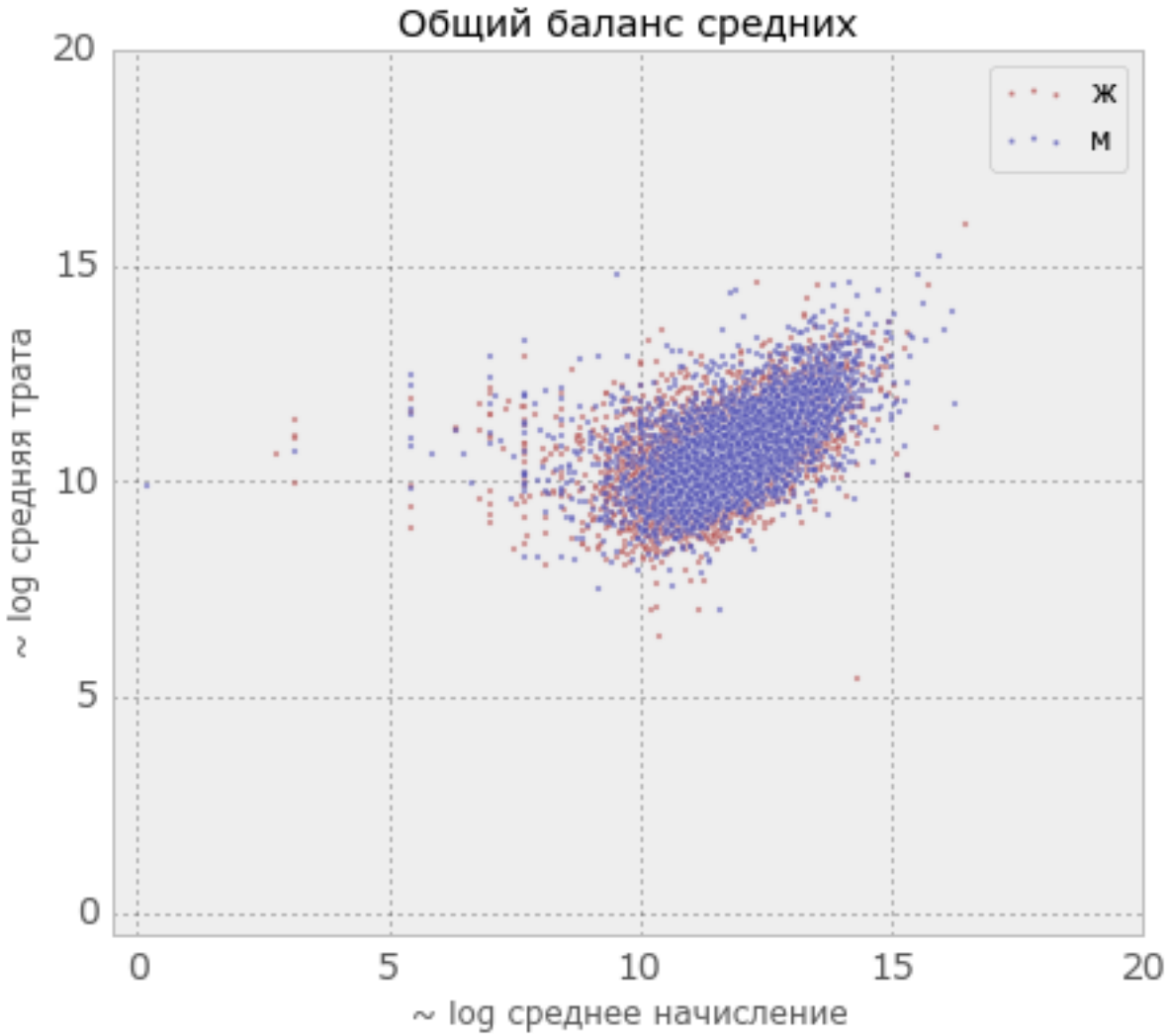
Что может быть виднее на «одномерных» картинках…
(тут плотности – уже отнормированные)

Суммы операций



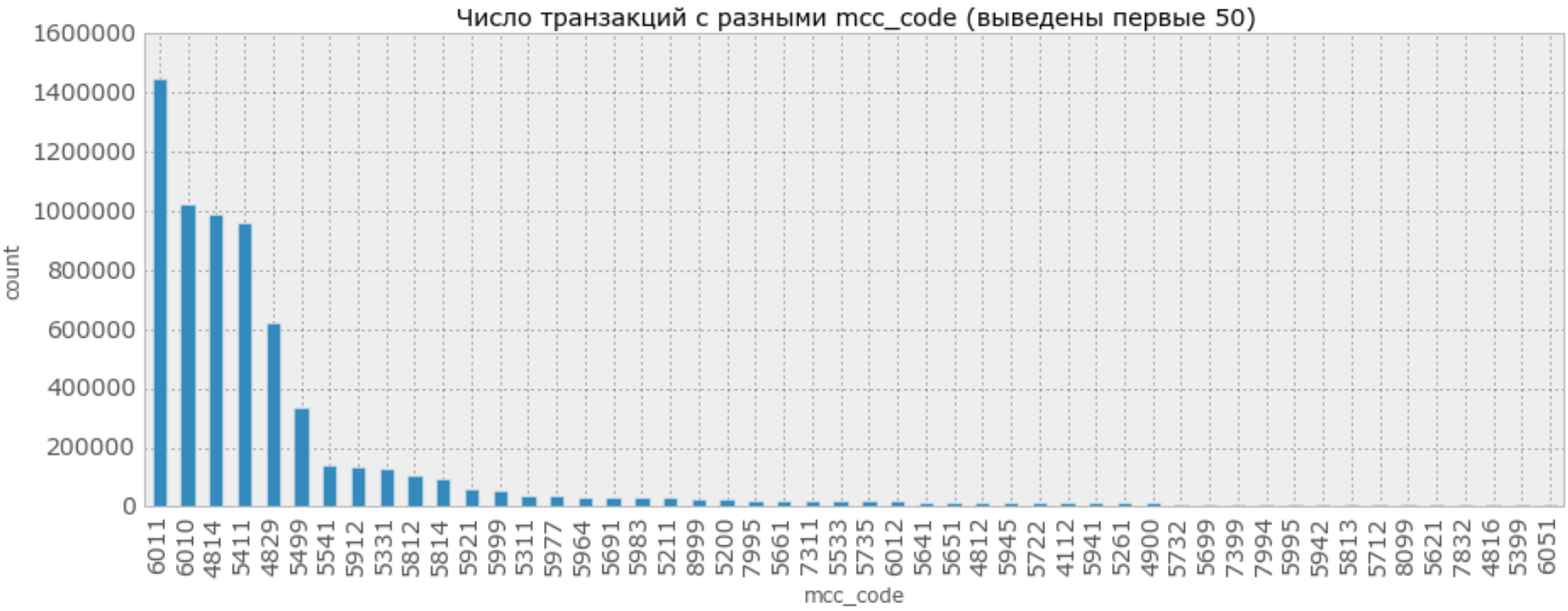
Мужчины снимают всё!

Суммы операций



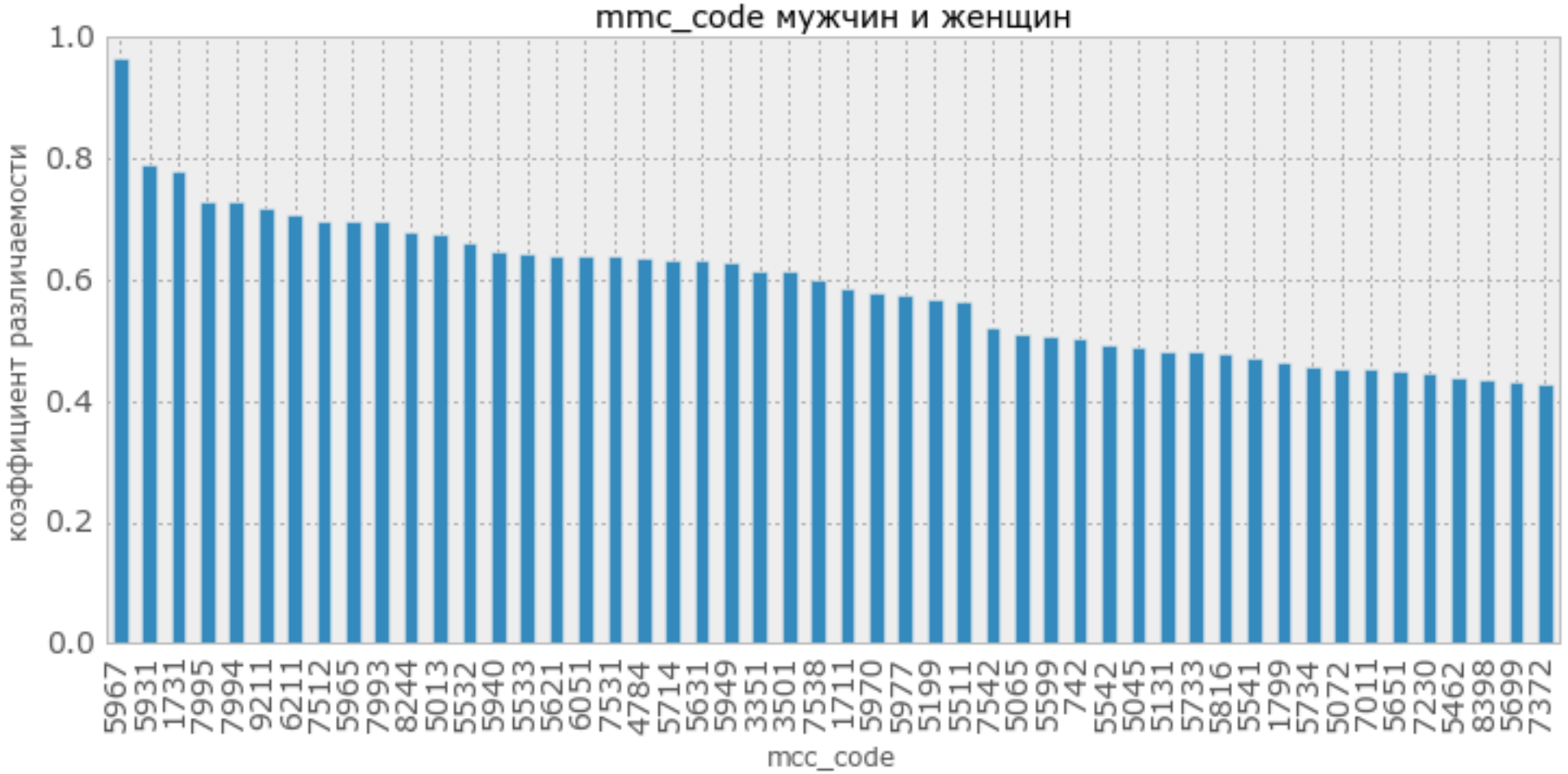
Что интересно?

Типы транзакций



Есть очень популярные mcc-коды

Гендерные покупки



Придумаем коэффициент гендерности (различаемости м/ж) кода (типичности для представителя одного пола)

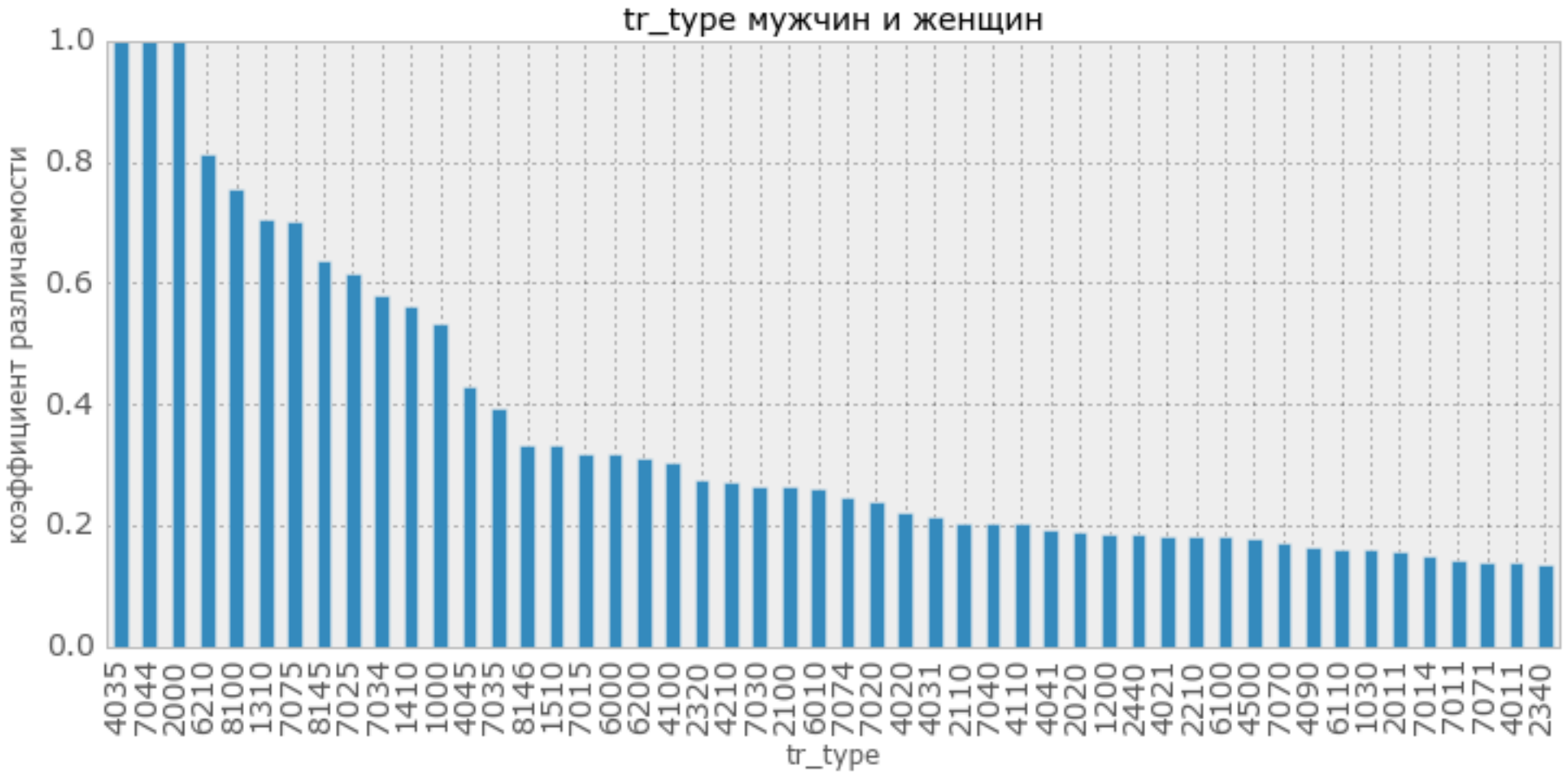
Гендерные покупки

mcc_code	ж	м	mcc_description	k
5967	5	289	Прямой маркетинг — входящий телемаркетинг	0.965986
5931	335	39	Магазины second hand, магазины б/у товаров, ко...	0.791444
1731	8	65	Подрядчики по электричеству	0.780822
7995	2431	15650	Транзакции по азартным играм	0.731099
7994	1164	7404	Галереи/учреждения видеоигр	0.728291
9211	43	7	Судовые выплаты, включая алименты и детскую по...	0.720000
6211	133	776	Ценные бумаги: брокеры/дилеры	0.707371
7512	22	123	Прокат автомобилей	0.696552
5965	106	19	Прямой маркетинг — комбинированный каталог и т...	0.696000
7993	106	591	Принадлежности для видеоигр	0.695839

Это очень хорошие гендерные признаки!

Типы транзакций

Типы транзакций



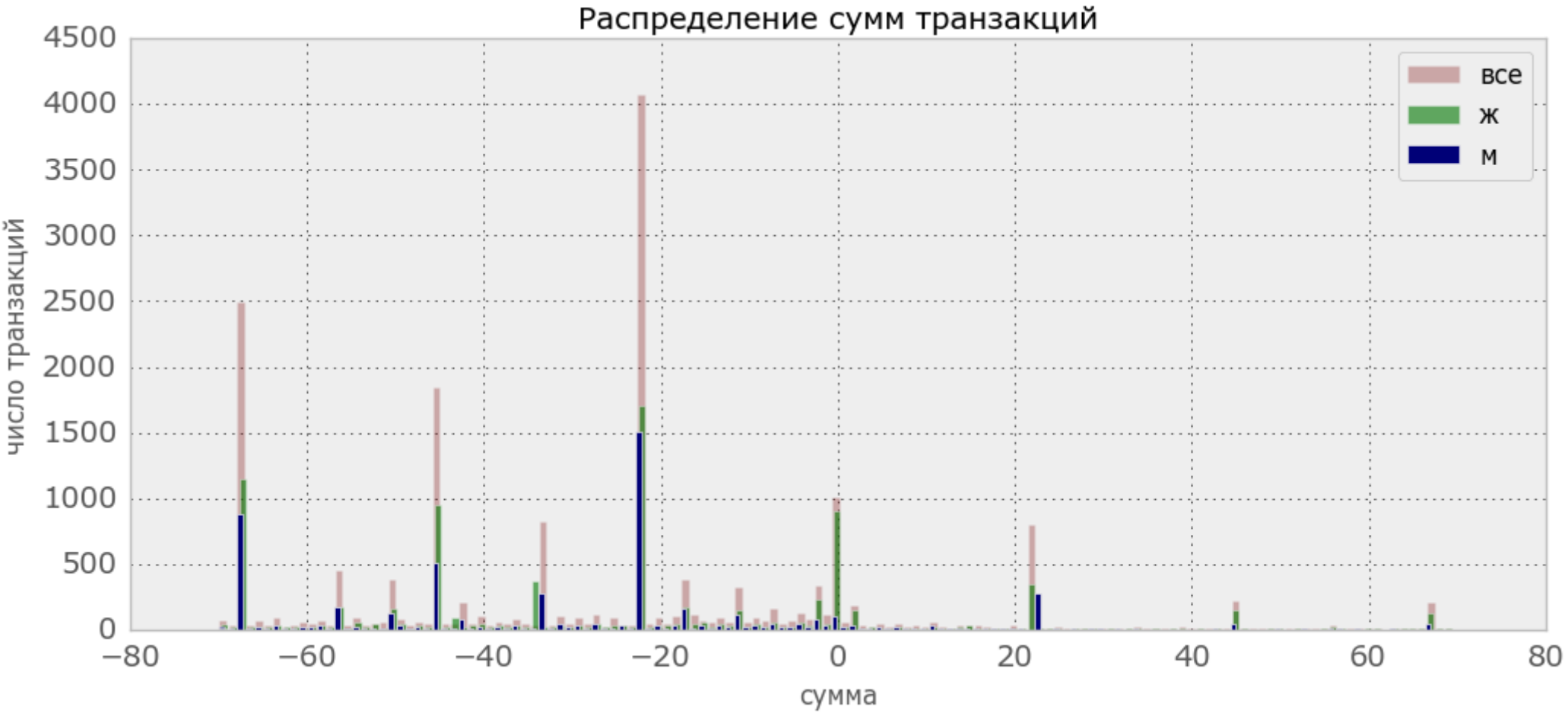
Но тут много редких...

Гендерные покупки

tr_type	ж	м	tr_type_description	k
6210	39	377	Возврат покупки.POS Зарубеж. ТУ	0.812500
8100	107	15	Списание после проведения претензионной работы	0.754098
1310	130	749	н/д	0.704209
7075	58	332	Перевод с карты на карту в овердрафте через Мо...	0.702564
7034	518	138	Перевод на карту/ с карты через АТМ (без взима...	0.579268
1410	240	854	н/д	0.561243
7035	52	119	Перевод на карту/ с карты через АТМ (со взиман...	0.391813

Предварительно убрали редкие tr_type

Суммы транзакций

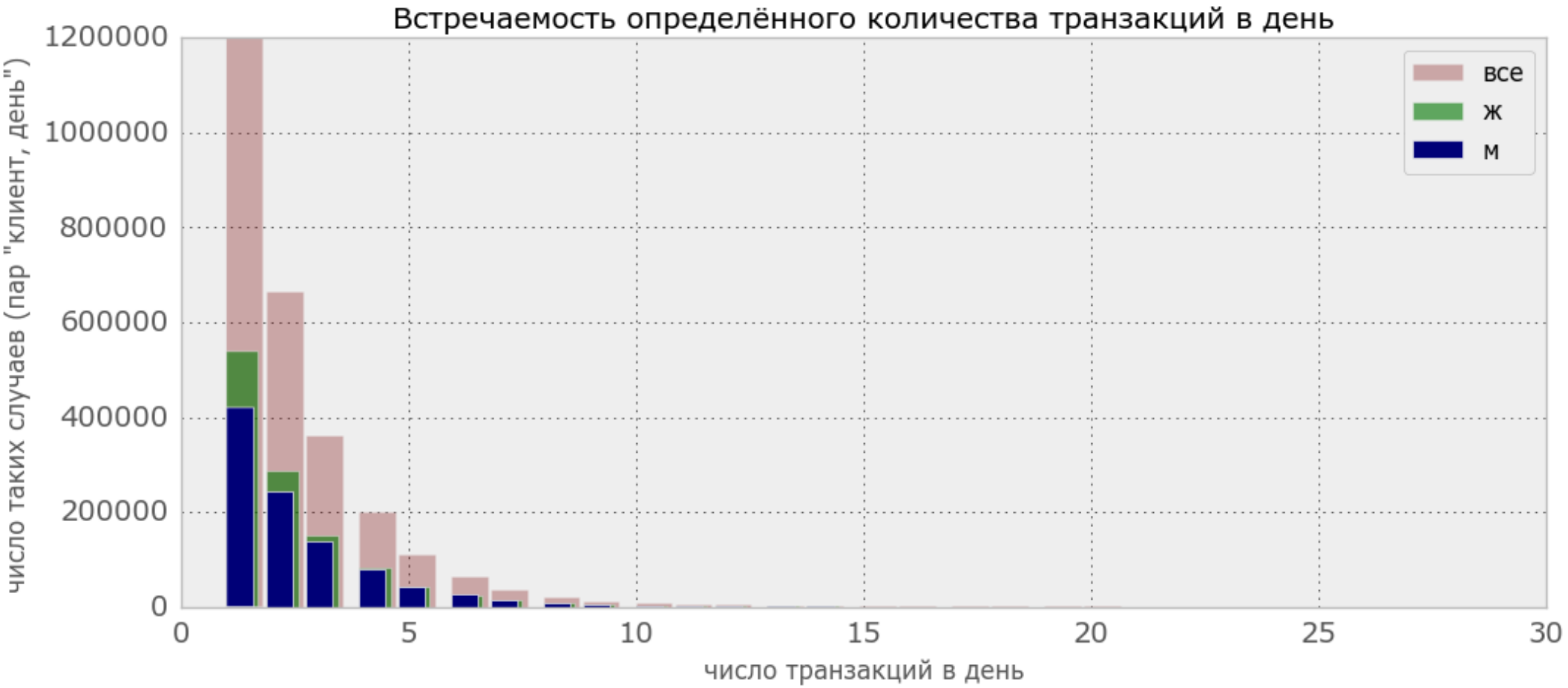


Только маленькие суммы

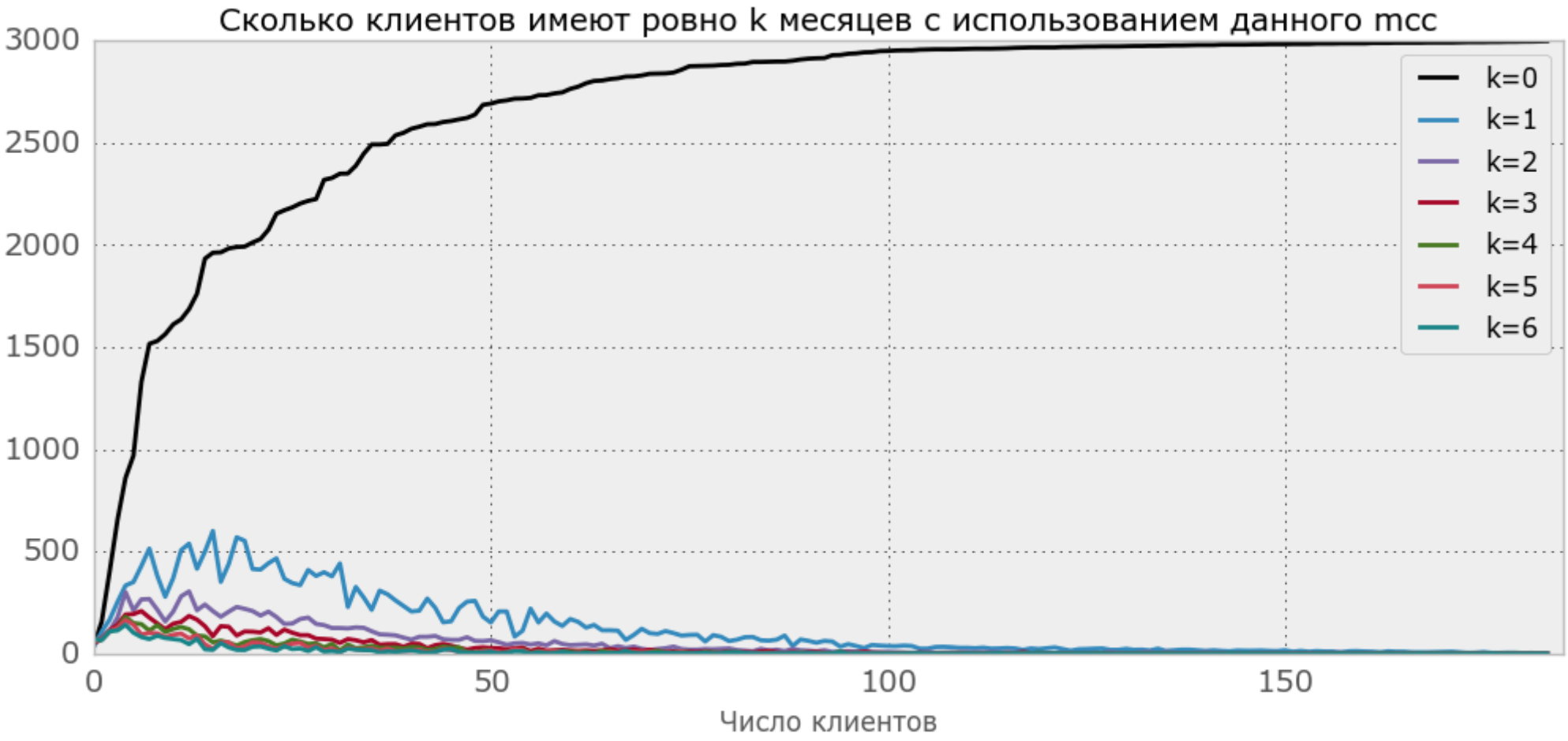
Понятно, что были изменены…

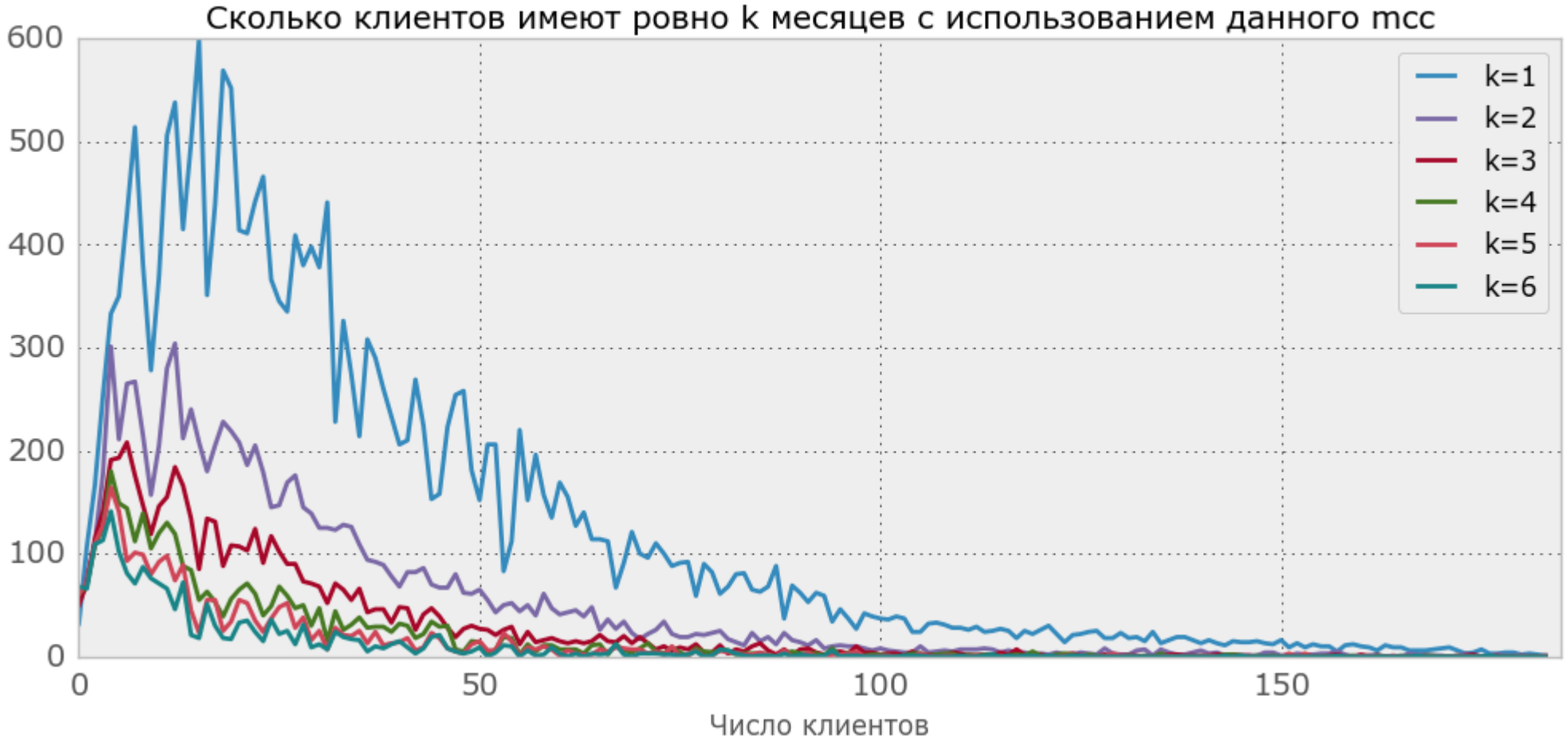
Неужели есть «Чисто женские суммы трат»?!

Сколько в день происходит транзакций



Редкие mcc


Сколько клиентов имеют ровно k месяцев с использованием данного mcc
3000
2500
2000
1500
1000
500
0
0
50
100
150
Число клиентов
k=0
k=1
k=2
k=3
k=4
k=5
k=6


Сколько клиентов имеют ровно k месяцев с использованием данного mcc
600
500
400
300
200
100
0
0
50
100
150
Число клиентов
k=1
k=2
k=3
k=4
k=5
k=6