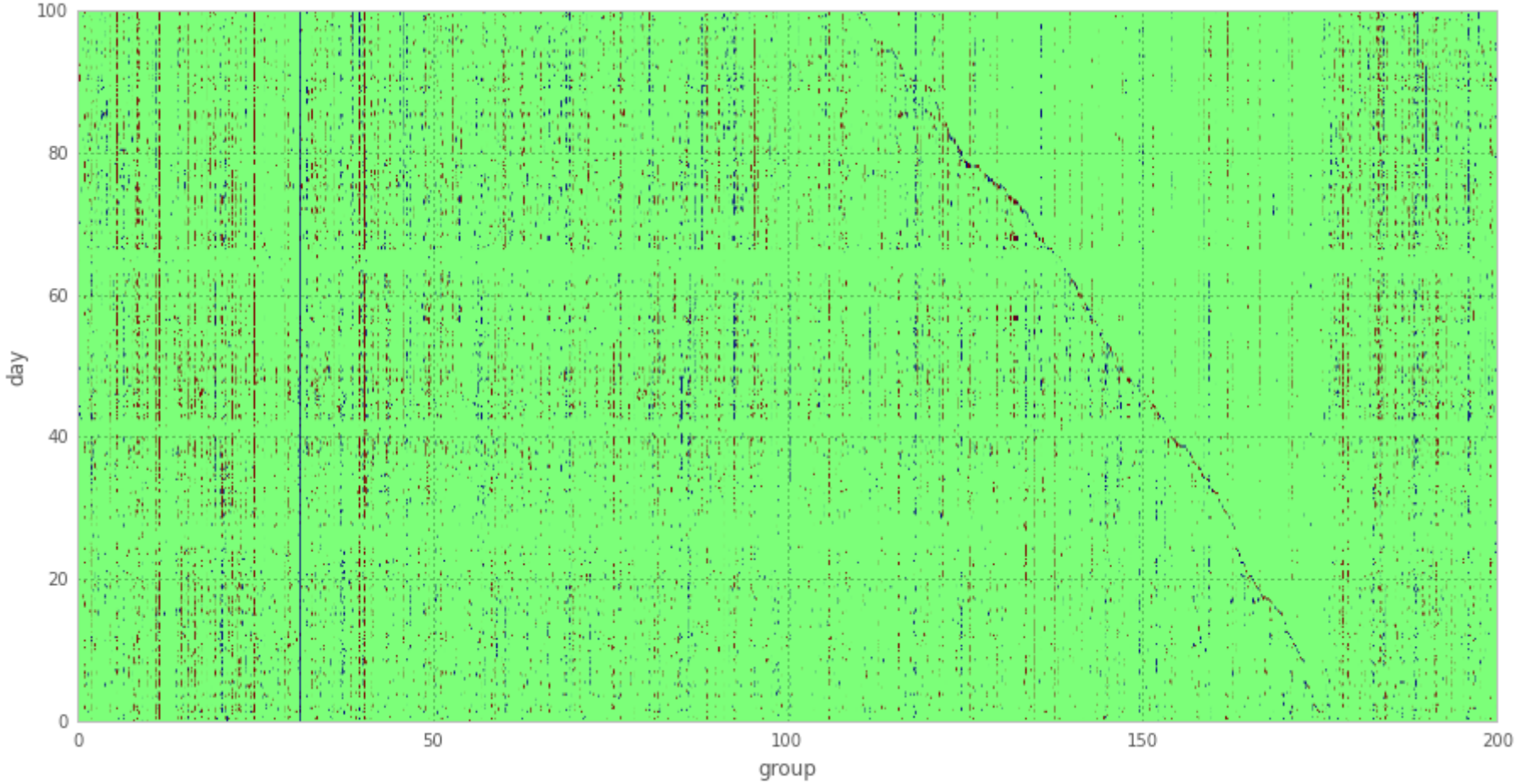
Прикладные задачи анализа данных

Case: ВИЗУАЛИЗАЦИИ

Дьяконов А.Г.

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

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:

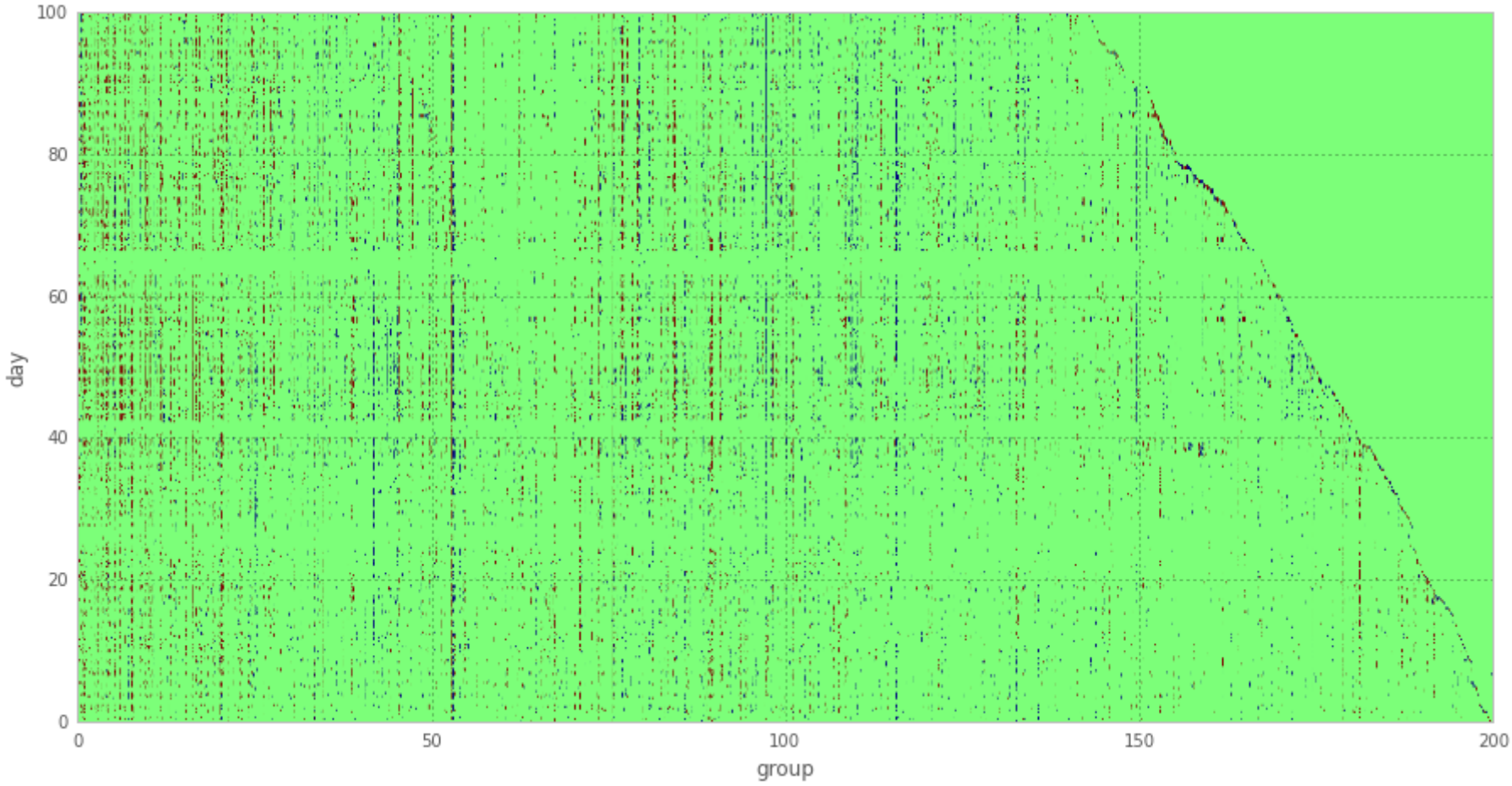
```
group_date2.columns[:10]

'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…

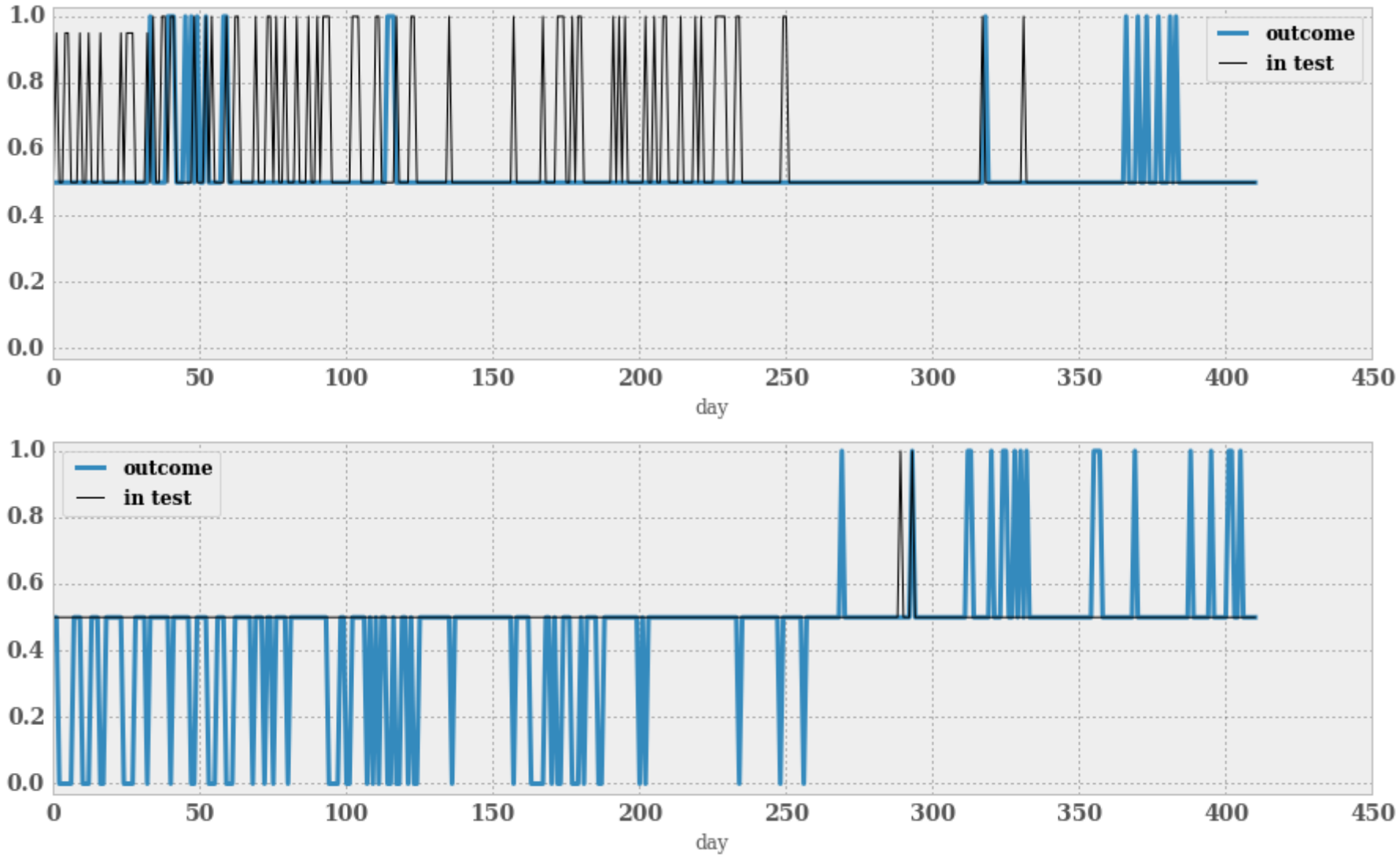
```
data_train.group_1 = data_train.group_1.map(lambda x: int(x[6:]))
```

Визуализация данных (RedHat)



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

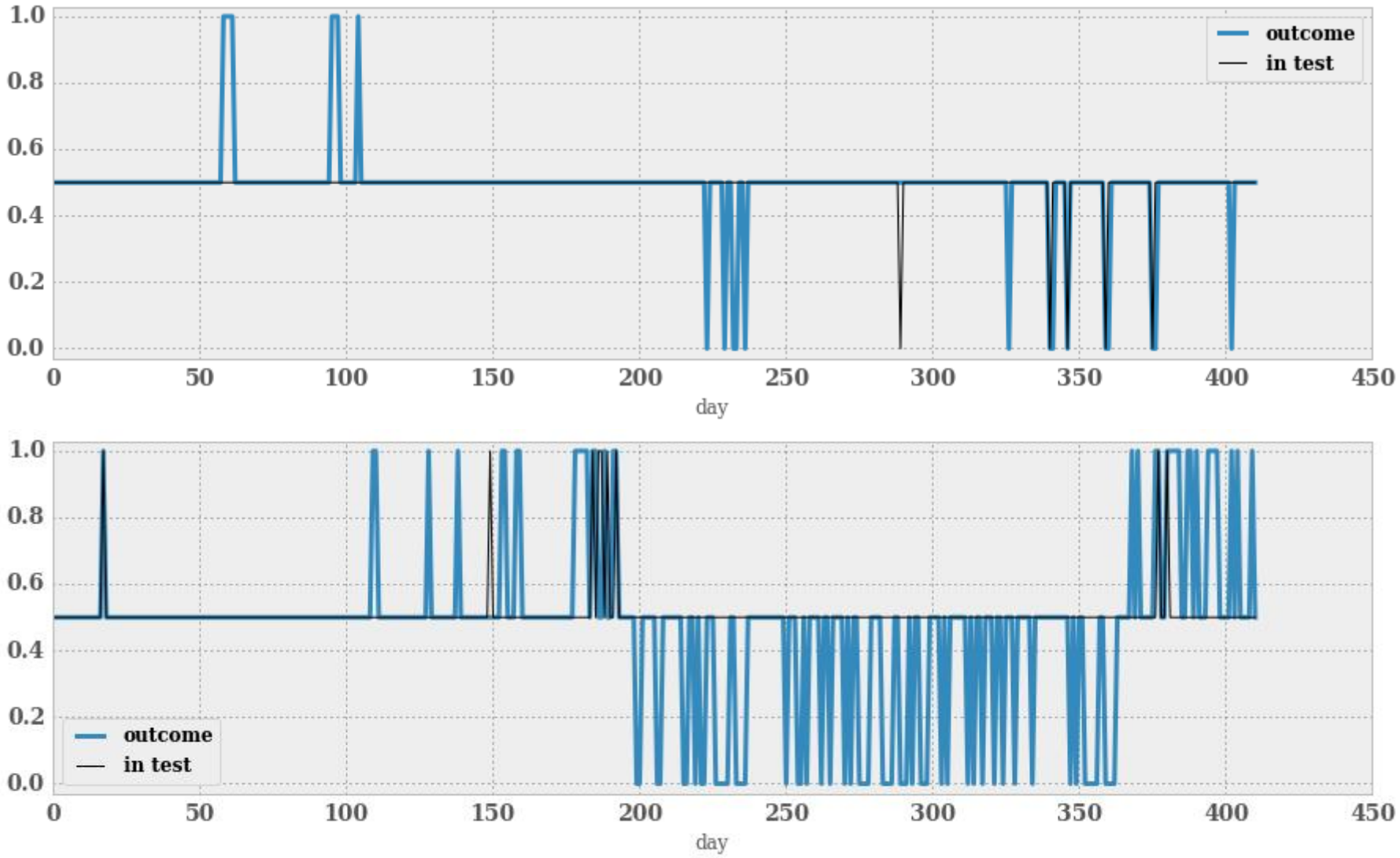
Задача «RedHat»



Как ведут себя представители групп по дням

Каждый график – для отдельной группы

Задача «RedHat»



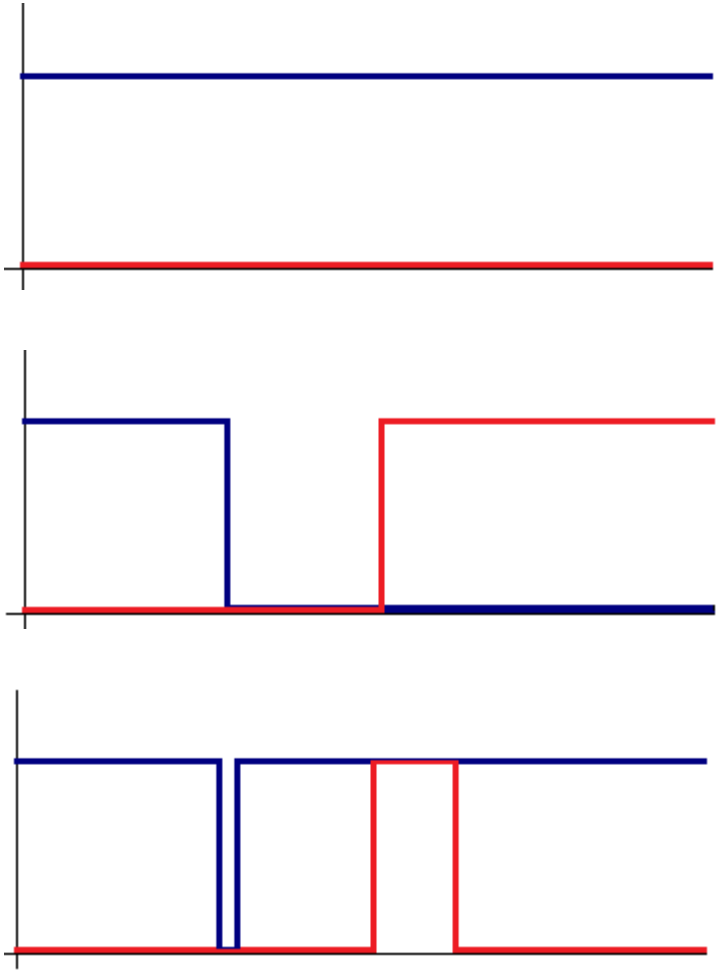
Как ведут себя представители групп по дням

Каждый график – для отдельной группы

Задача «RedHat»

Что видим?

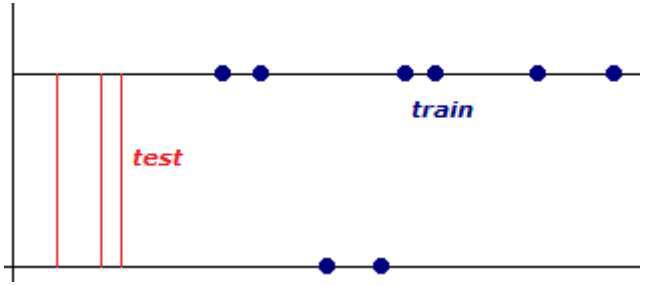
целевой признак кусочно-константный



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

Задача об оценке 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

Задача об оценке эффективности менеджера

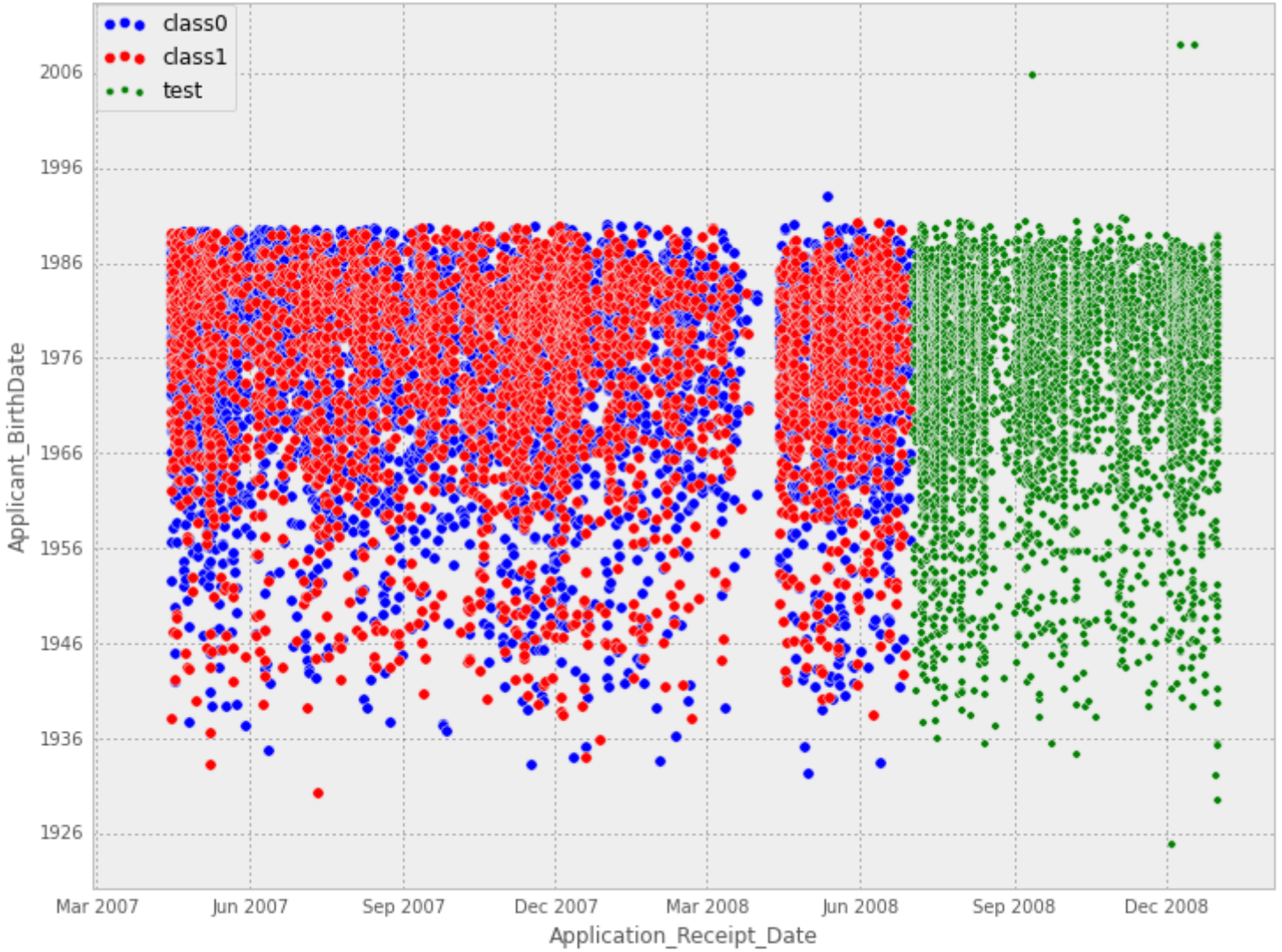
Смотрим данные – делаем гипотезы

	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

Задача об оценке эффективности менеджера

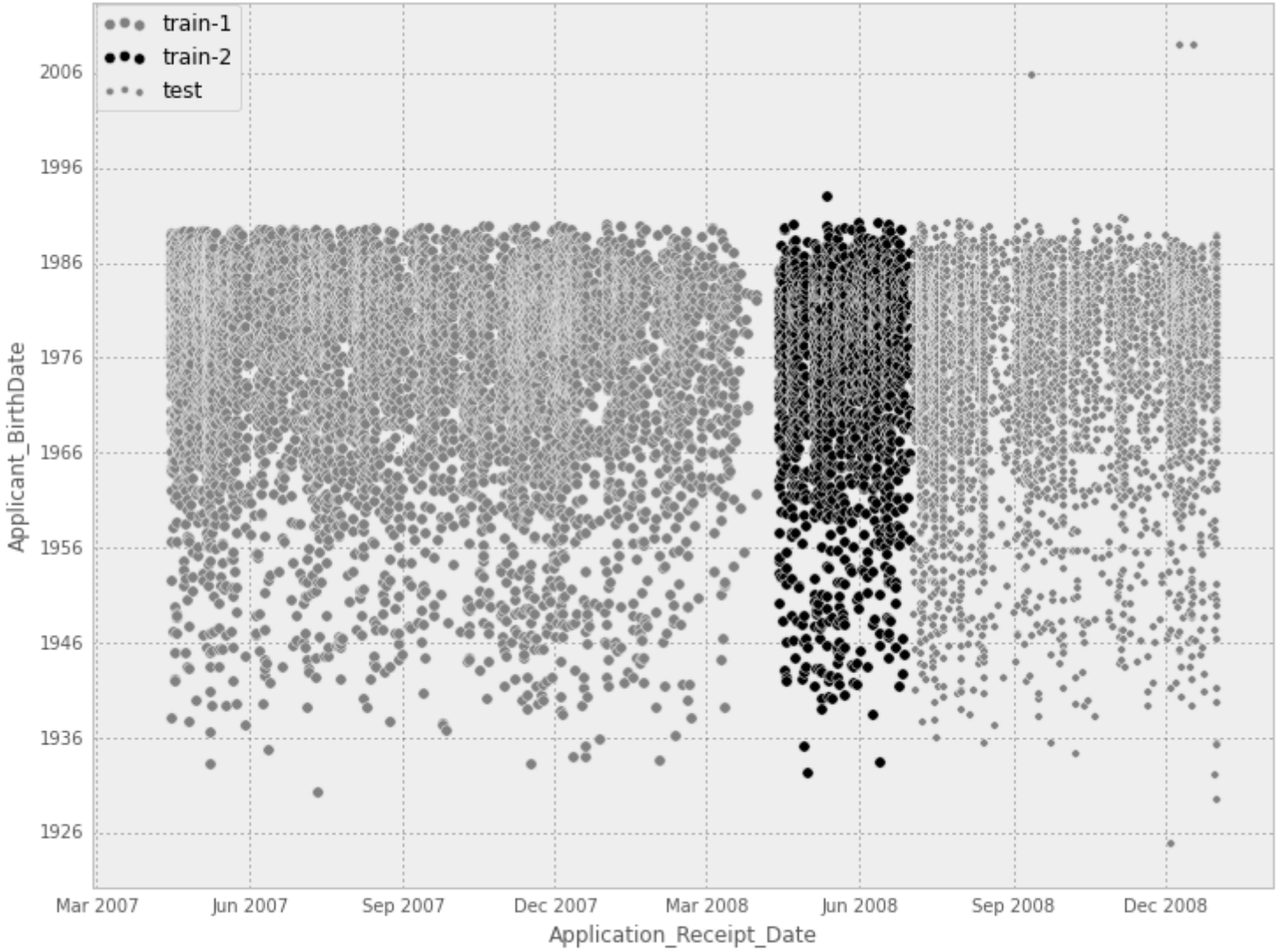
Признак «время сделки» по горизонтали



Что интересно?

Задача об оценке эффективности менеджера

Признак «время сделки» по горизонтали



Начало нового фрагмента... как это использовать?

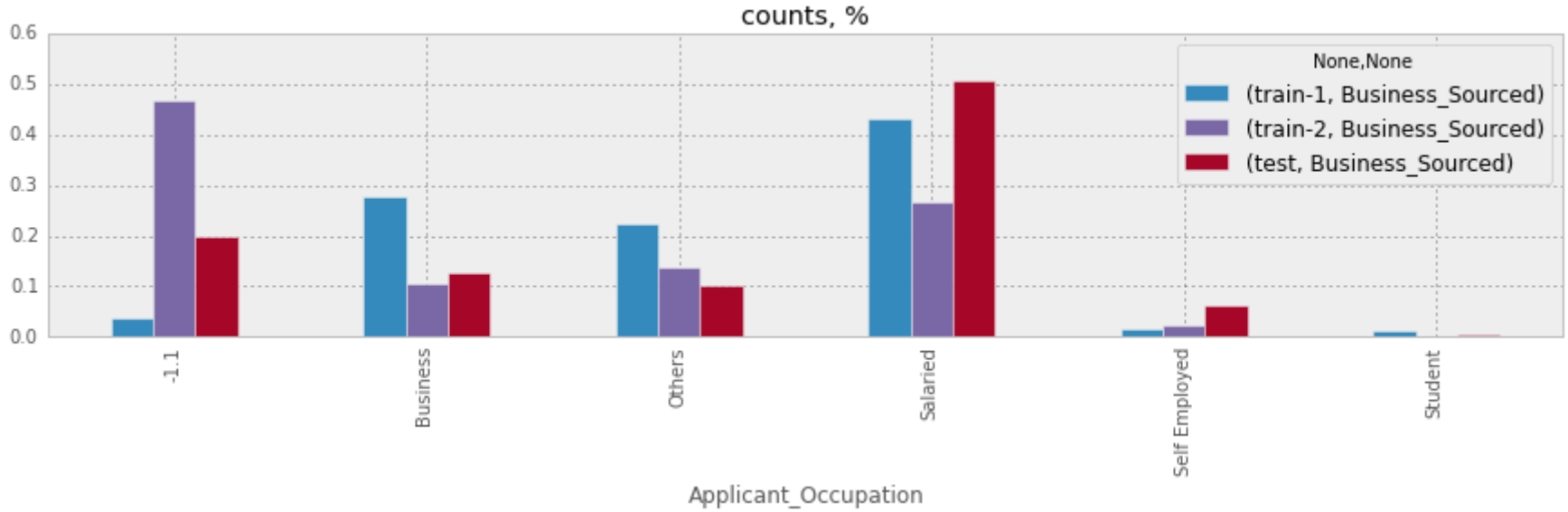
Задача об оценке эффективности менеджера

Если делать контроль CV – качество 0.65 AUC ROC

Если контроль – последний кусок обучения – 0.55 AUC ROC

Теперь ясно почему!

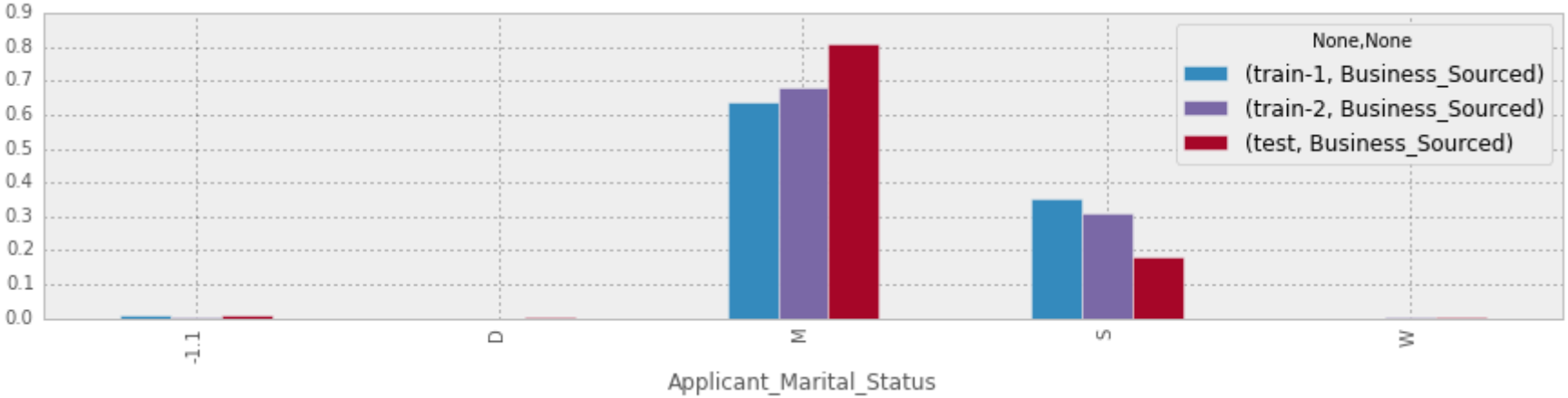
Распределение рода занятий в разных кусках



Задача об оценке эффективности менеджера

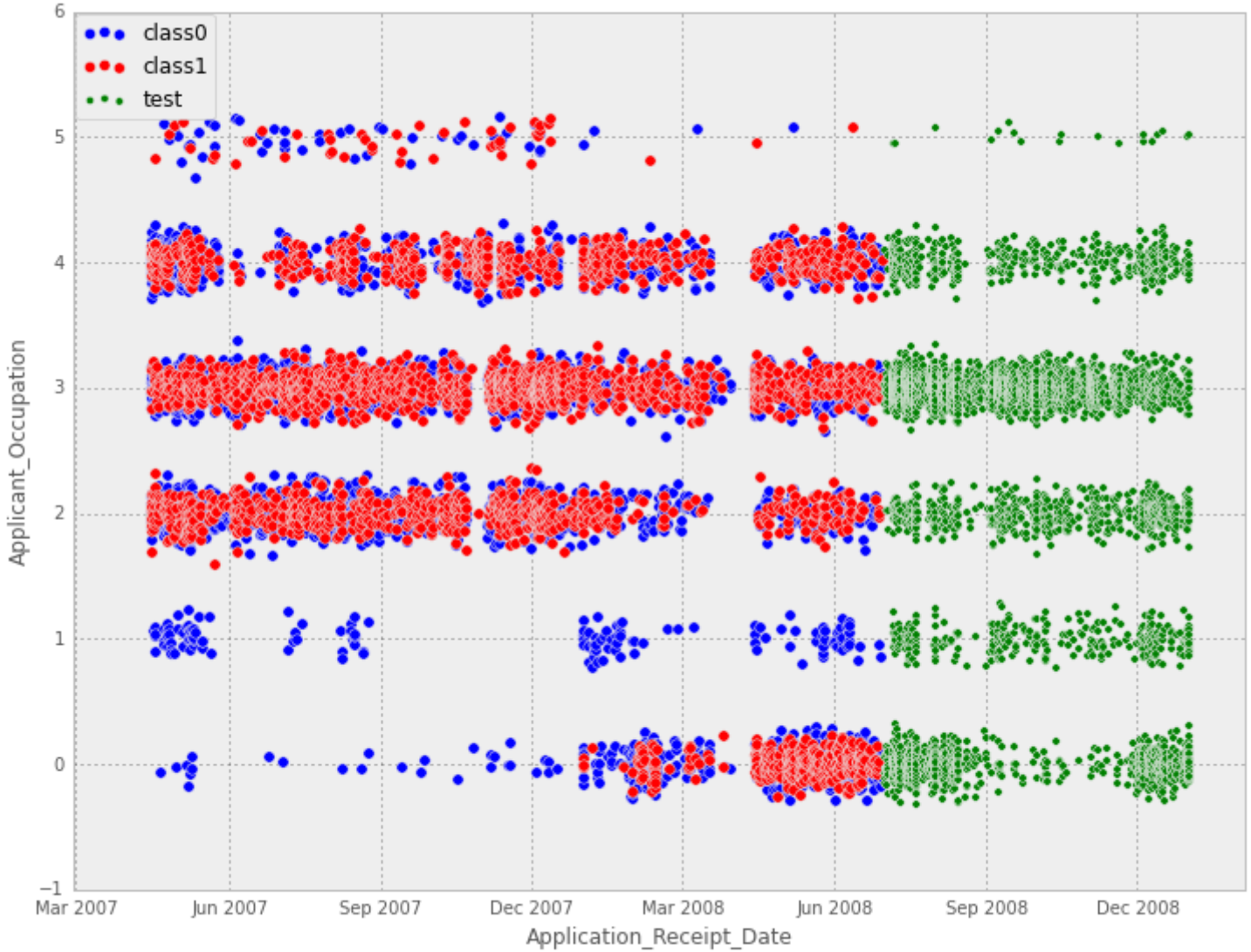
Распределение пола в разных кусках

Распределение семейного положения



Задача об оценке эффективности менеджера

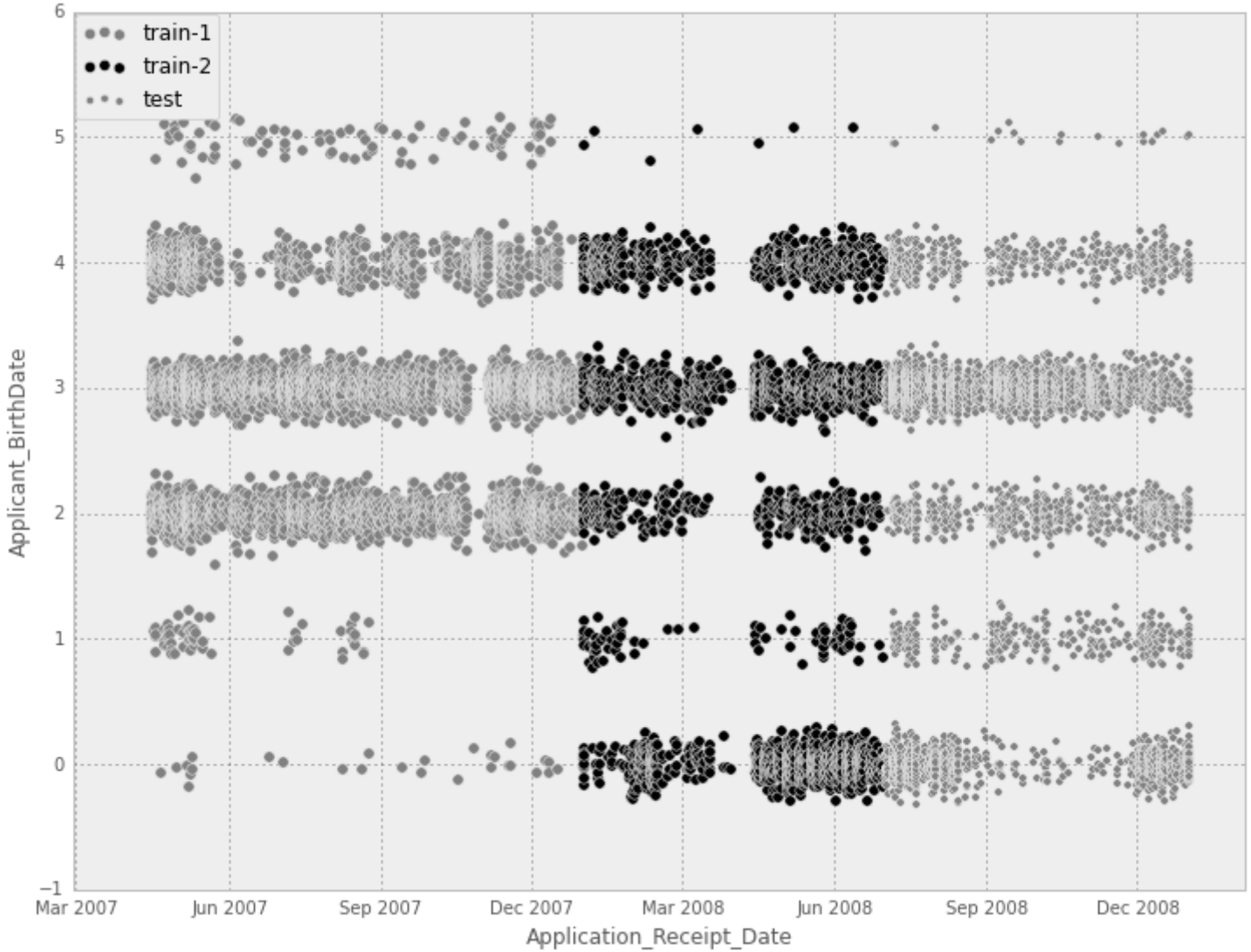
Изменение распределений признаков во времени (сделан jitter)




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

Задача об оценке эффективности менеджера

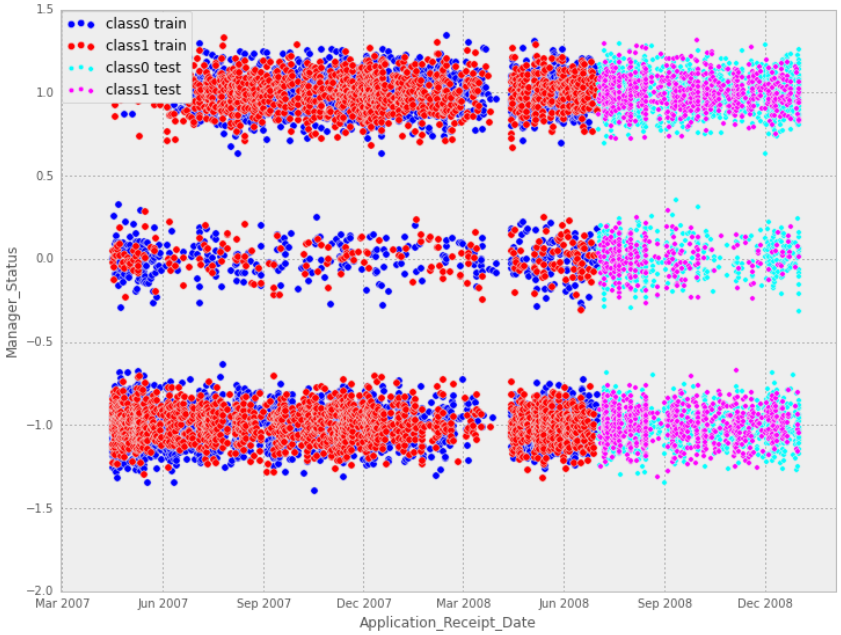
Изменение распределений признаков во времени



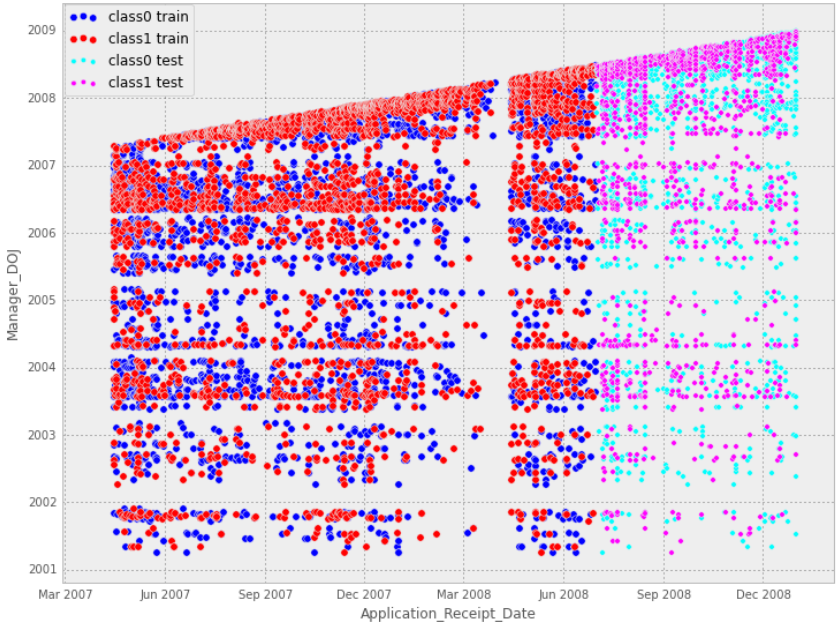
С 1 января 2008 года! По другим признакам подобного нет!!!

Задача об оценке эффективности менеджера

Статус менеджера (подсвечены ответы алгоритма)



Дата сделки / начало работы менеджера



Задача об оценке эффективности менеджера

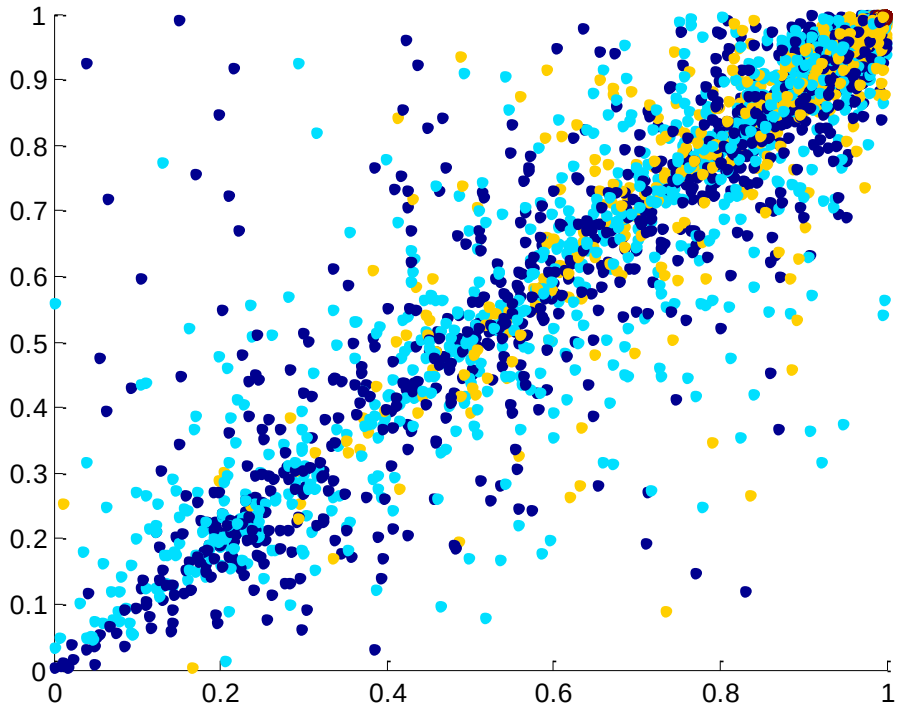
Интересный приём:

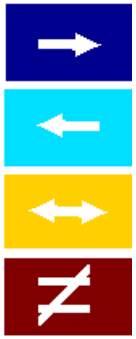
по train1 кодировать признаки, на train2 обучать…

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')
```

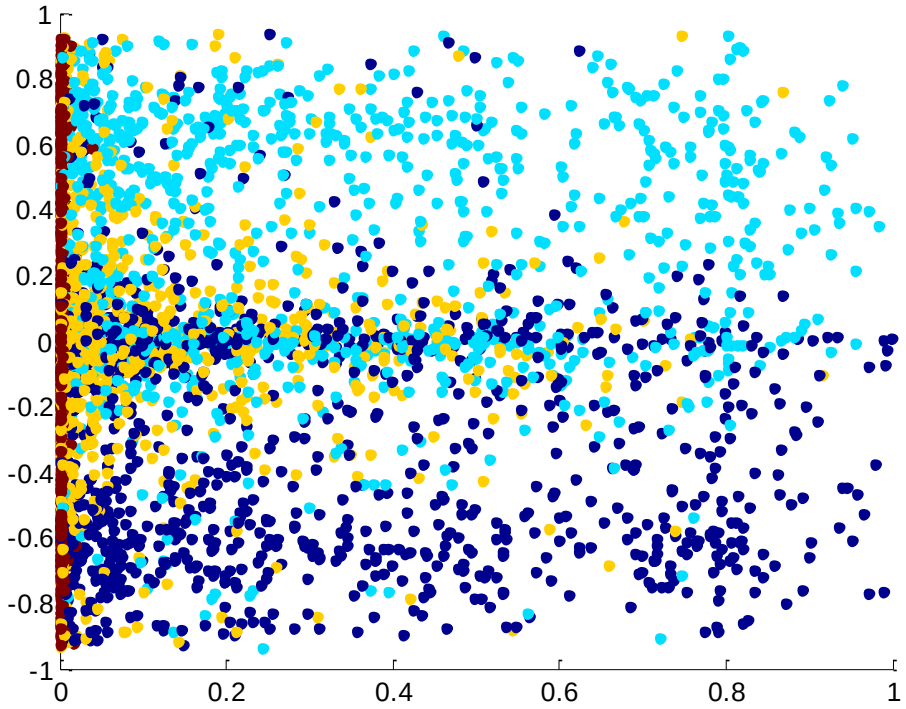




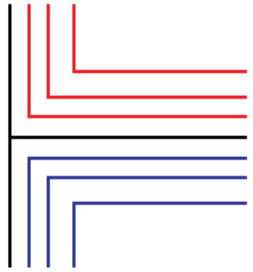
Кстати: хорошая задача – пример «новой науки»

Алгебраические выражения над признаками

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

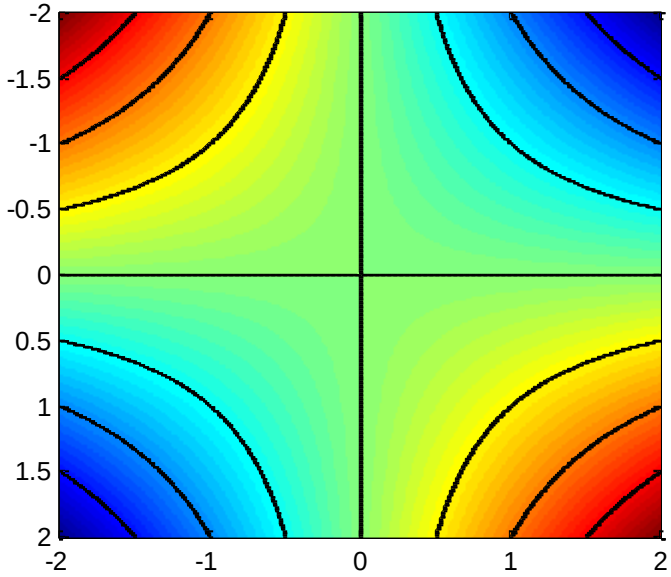


А теперь надо «уголками откусывать классы»:

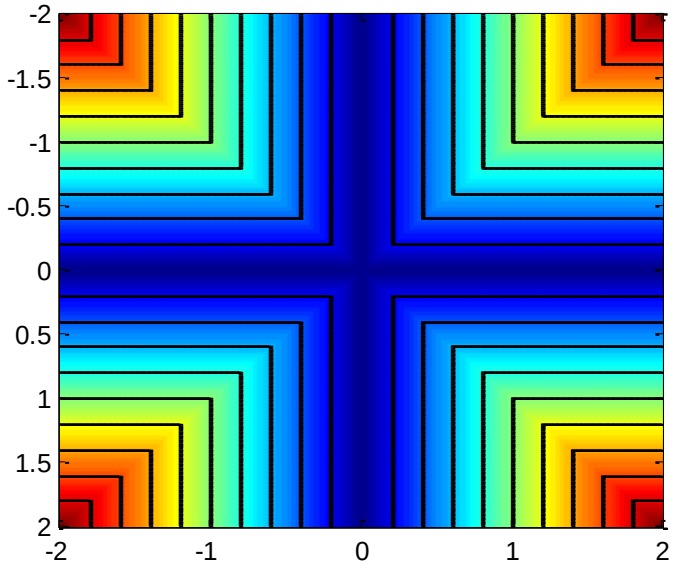


Какие функции «откусывают уголками»

$$z = y \cdot x$$

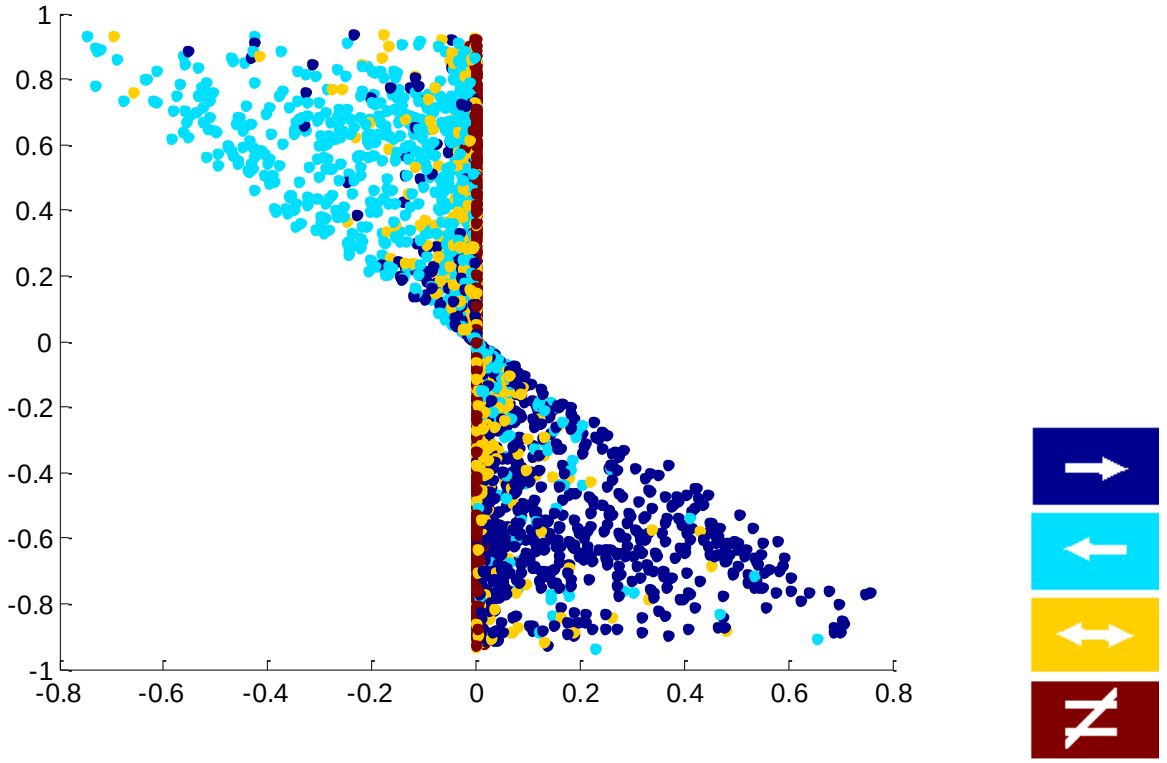


$$z = \min(|y|,|x|)$$



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')
```

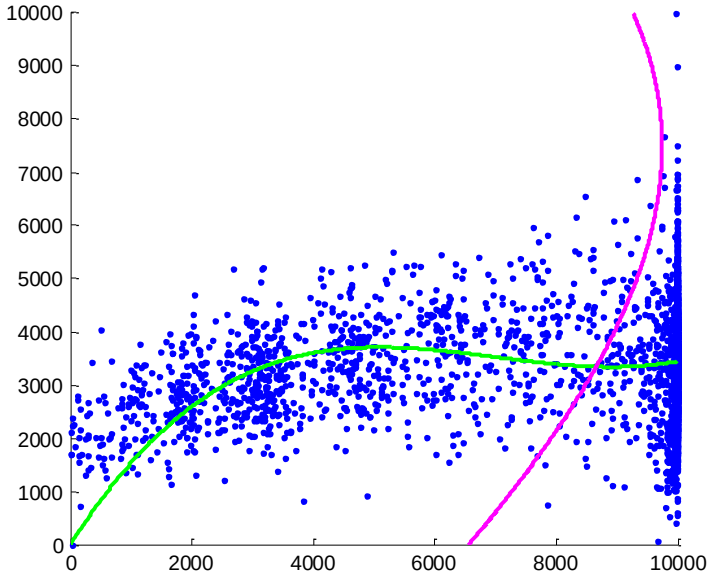


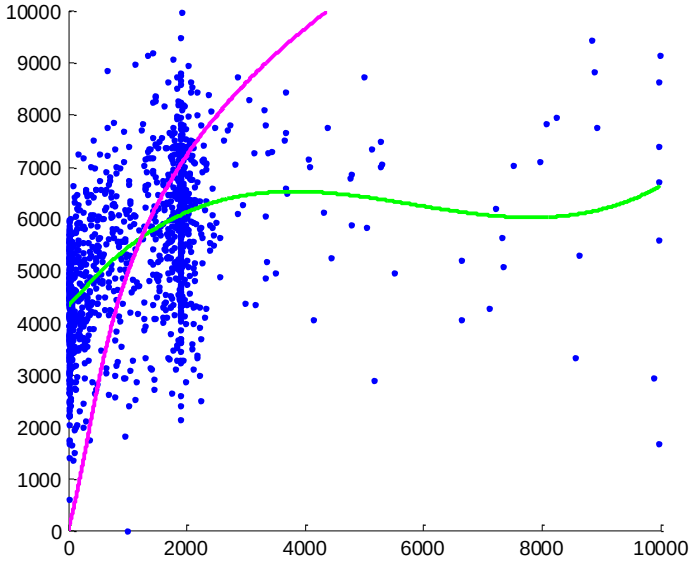
И здесь мы видим разделяемость синих и голубых!

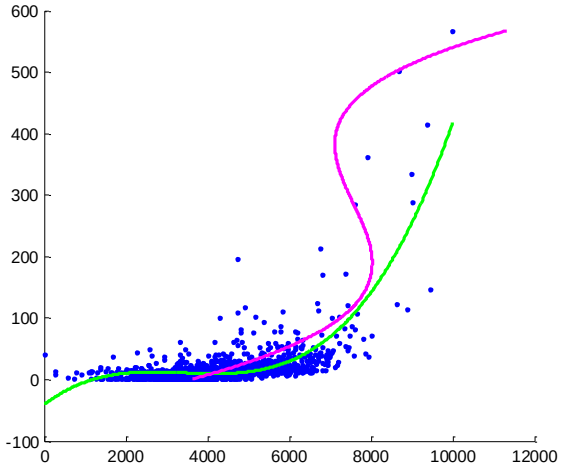
Получается алгоритм неплохого качества.

Ещё один приём: посмотреть как метод «работает»

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

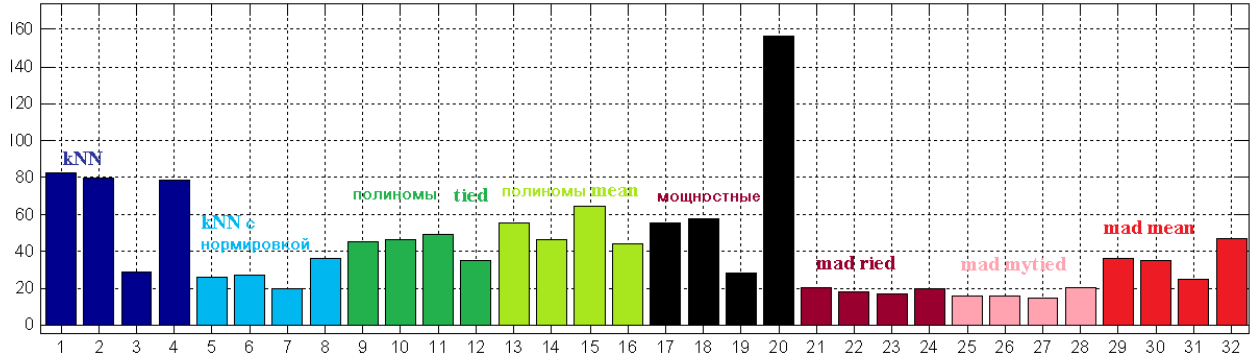




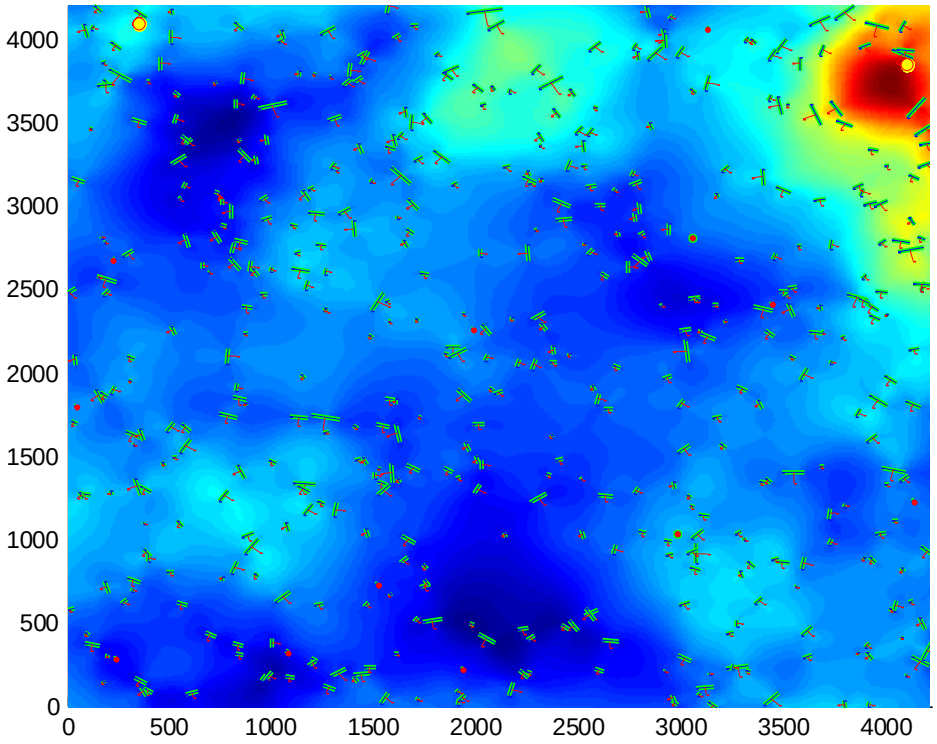


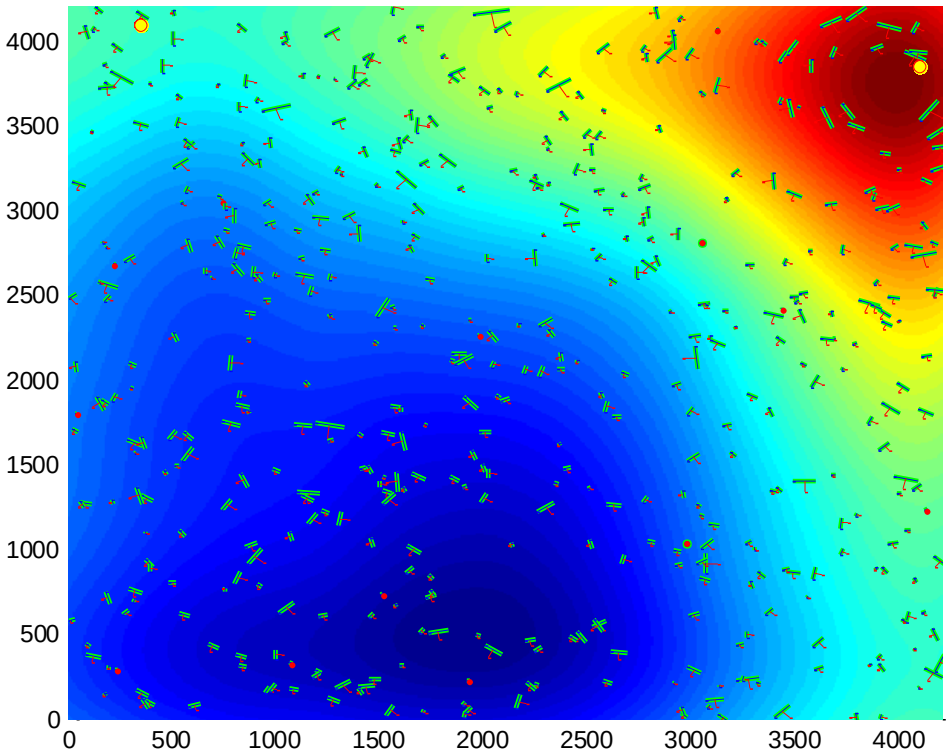
Ответы алгоритмов – как признаки

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



Задача про чёрные дыры



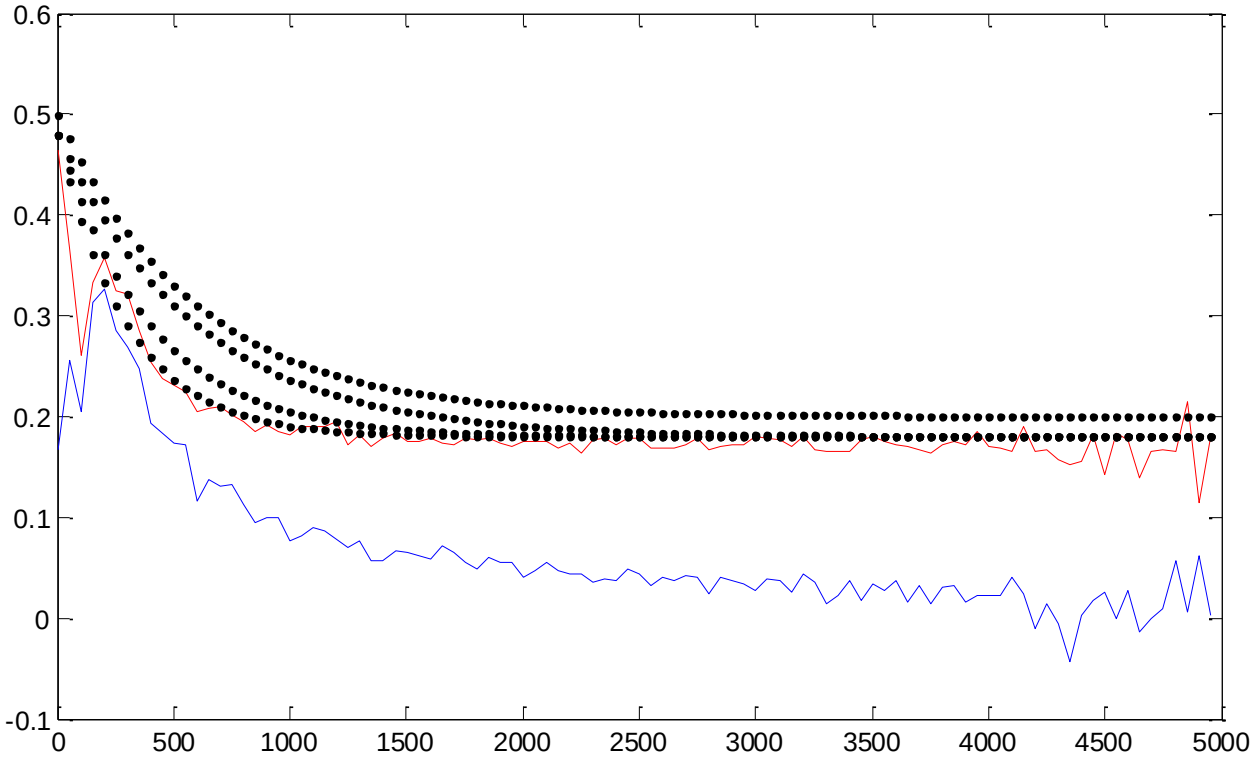


Какая связь между рисунками?

Ответ:

«Плотность» и её сглаженный аналог.

Средний профиль плотности(красный):



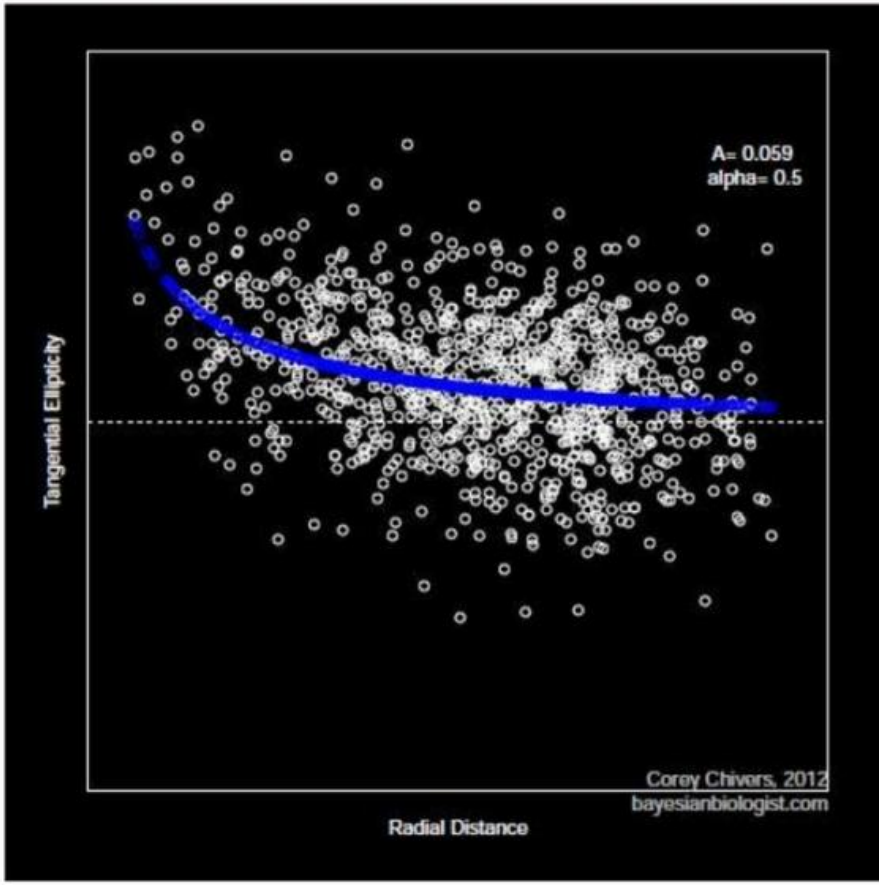
и методы его приближения

Решение Owen Zhang

Observing Dark Worlds competition

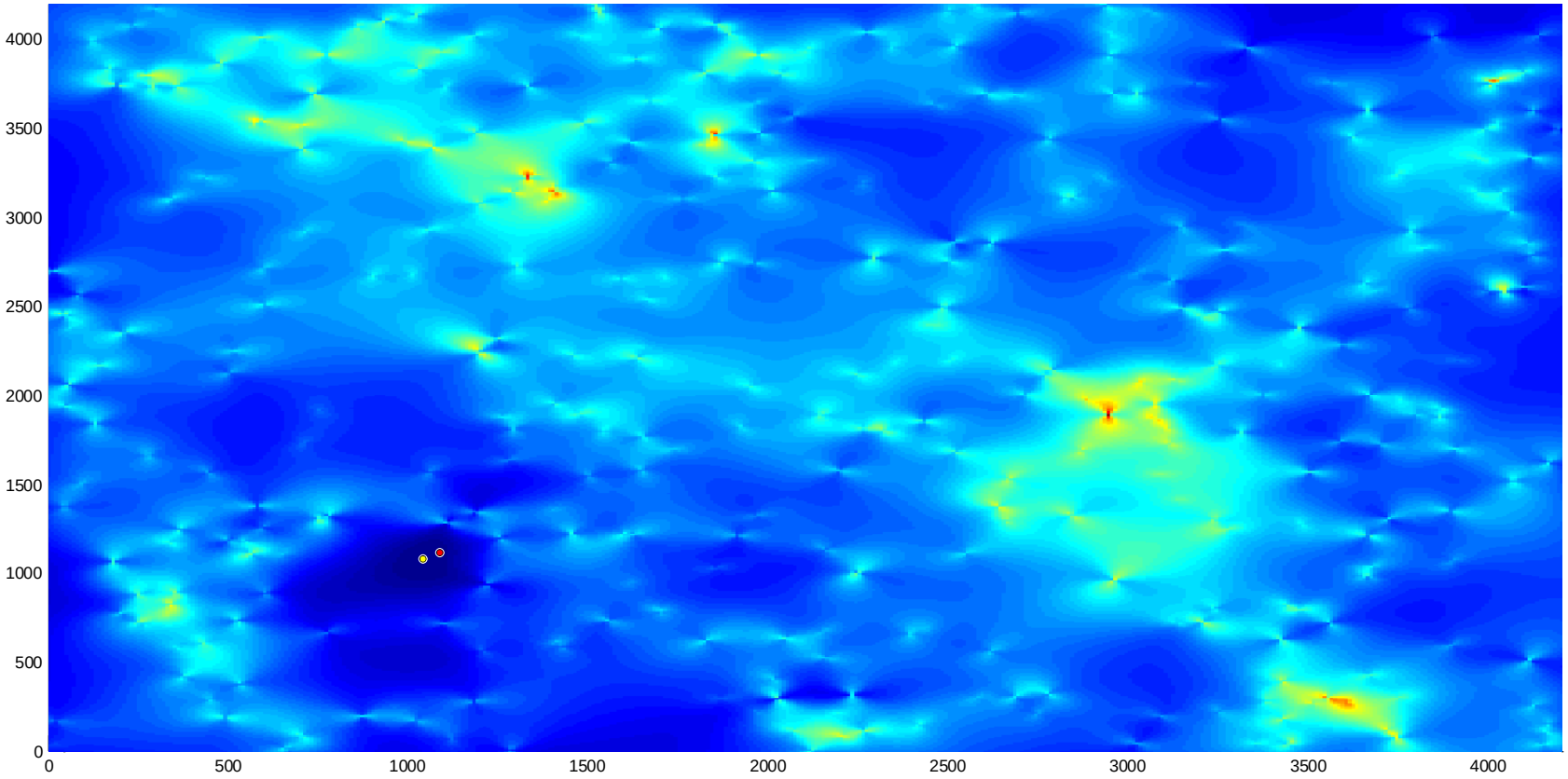
- Model P(Y|X):
  - Distortion is tangential to dark matter halo
  - Strength of the effect declines with 1/r
  - Strength of effect depends linearly on mass of halo

$$e_t \approx \frac{m}{r}$$



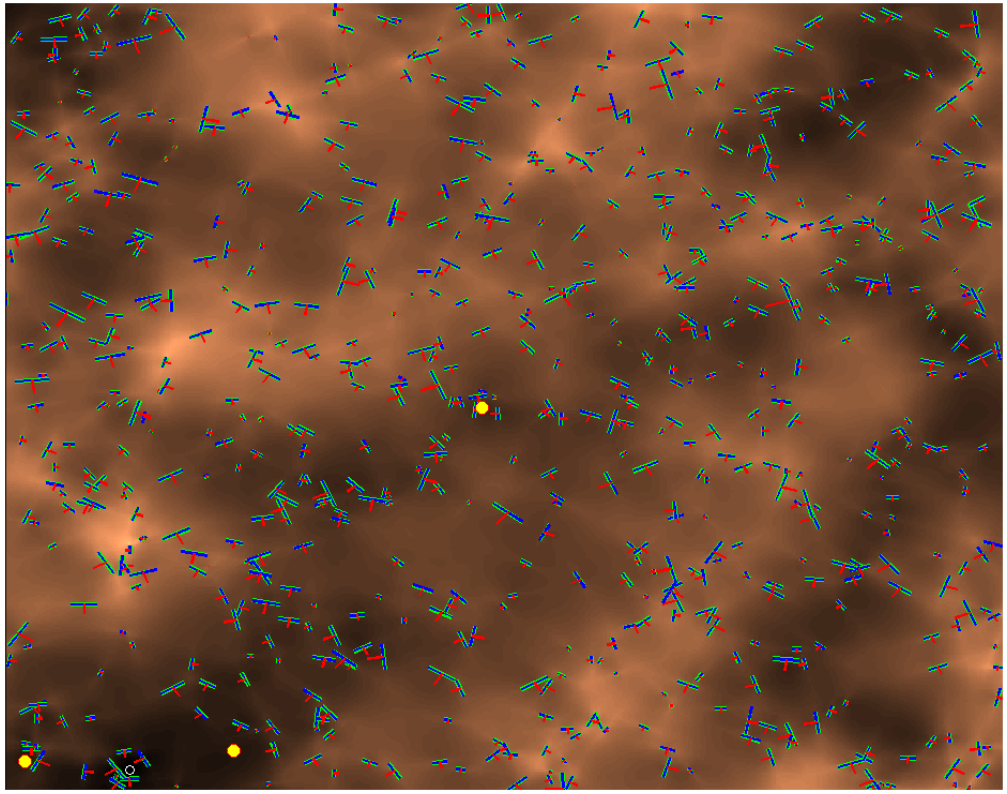
Также использовал визуализацию для создания модели

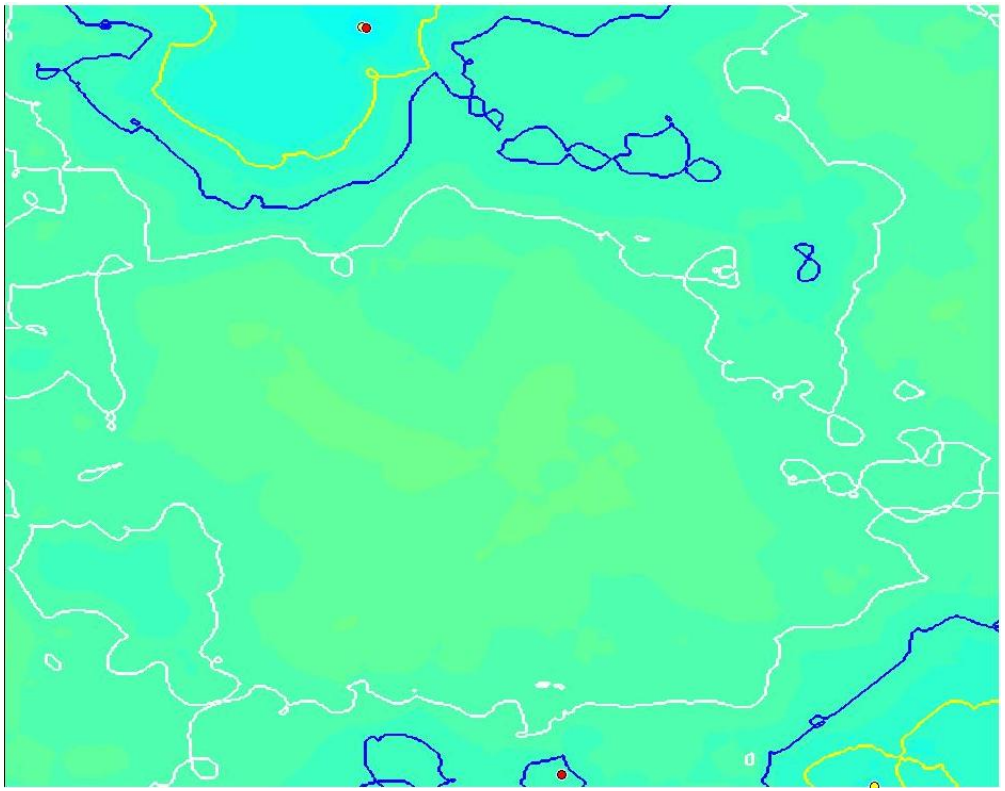
Другой способ:



разумно решать комбинацией двух

Трудности большого числа дыр:





переход к линиям уровня

Главное – выбор эффективной визуализации.

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

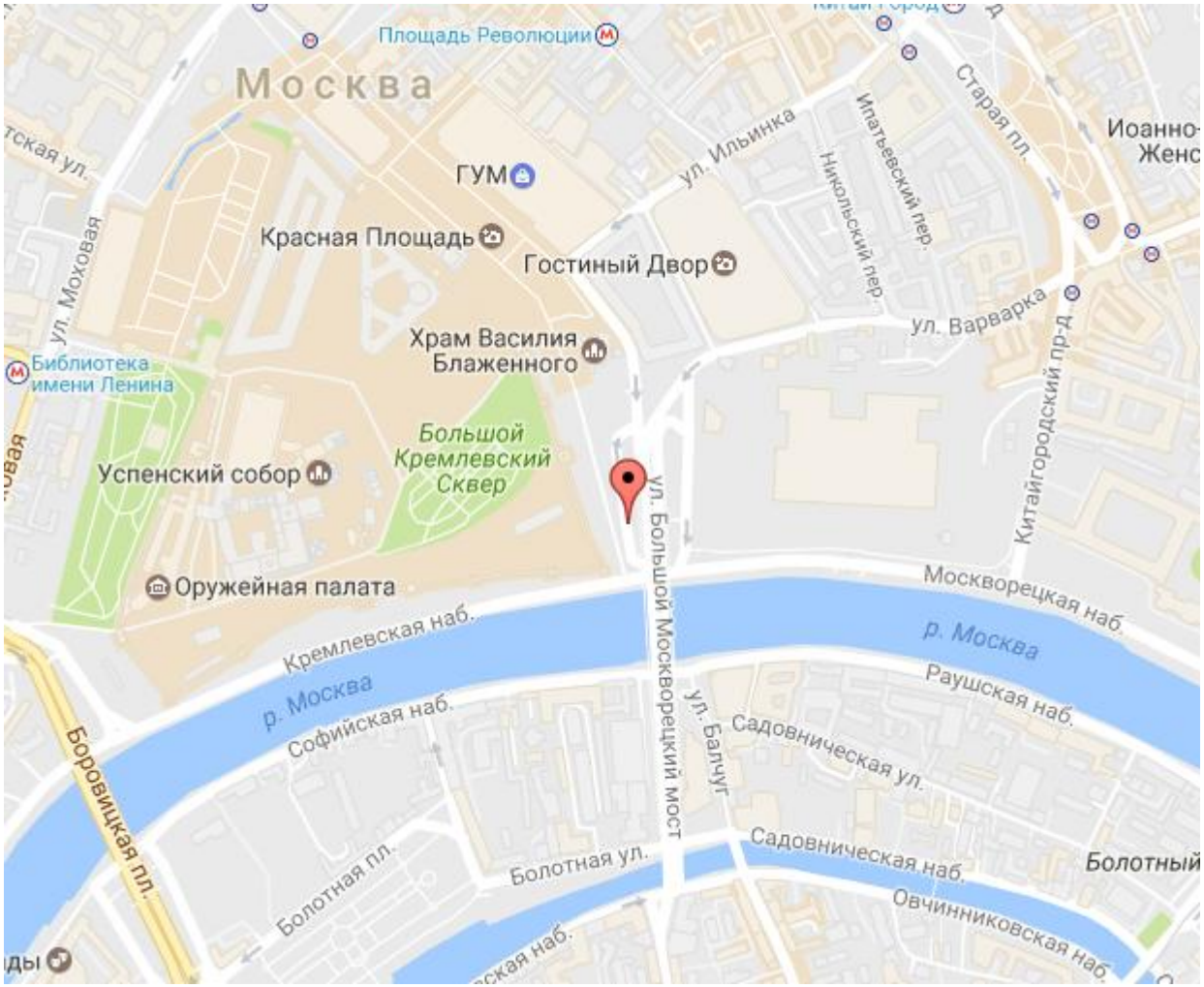
По какому принципу упорядочены данные?

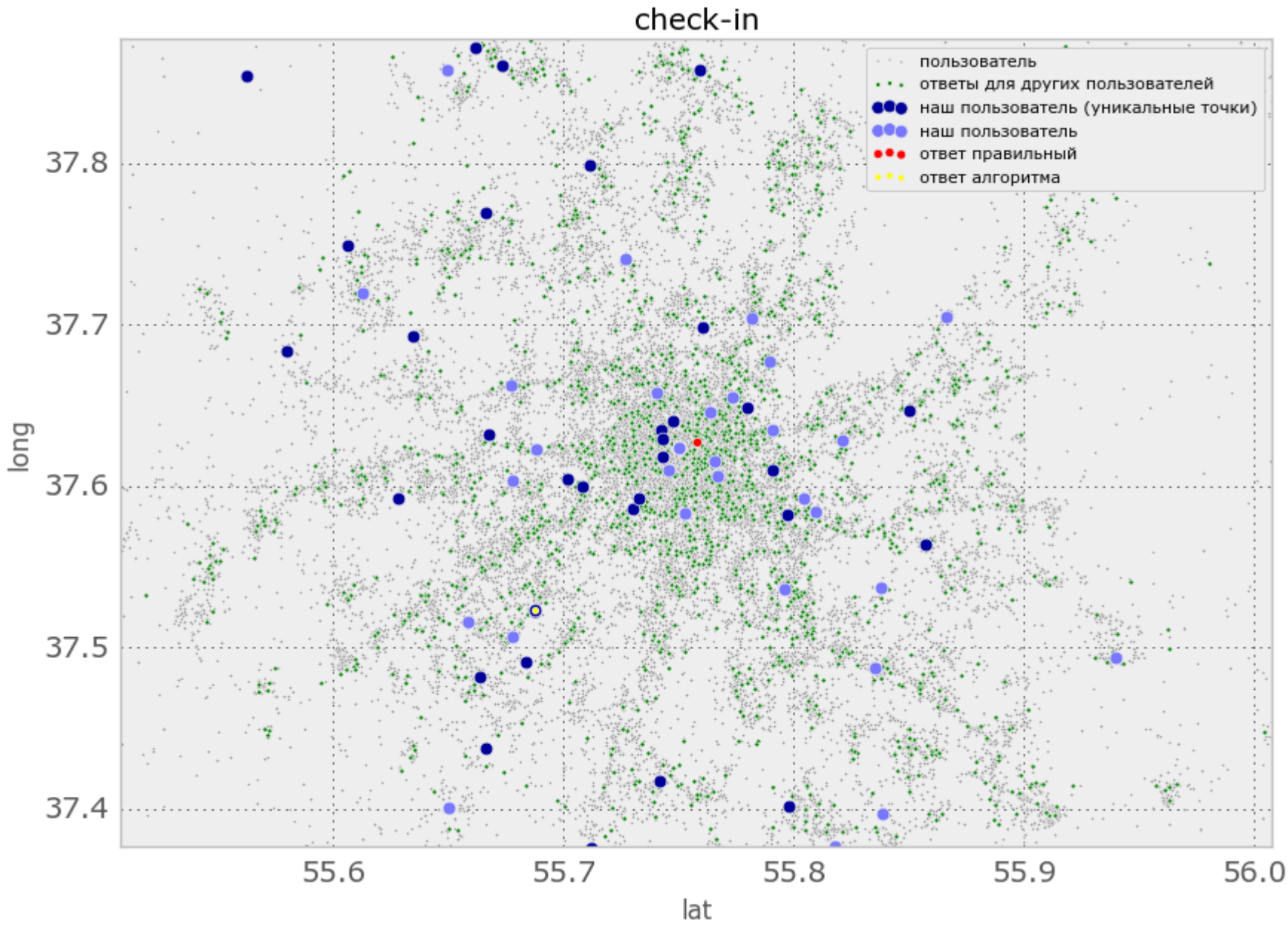
По дням… просто даты настоящих дней забиты «2017-01-15»

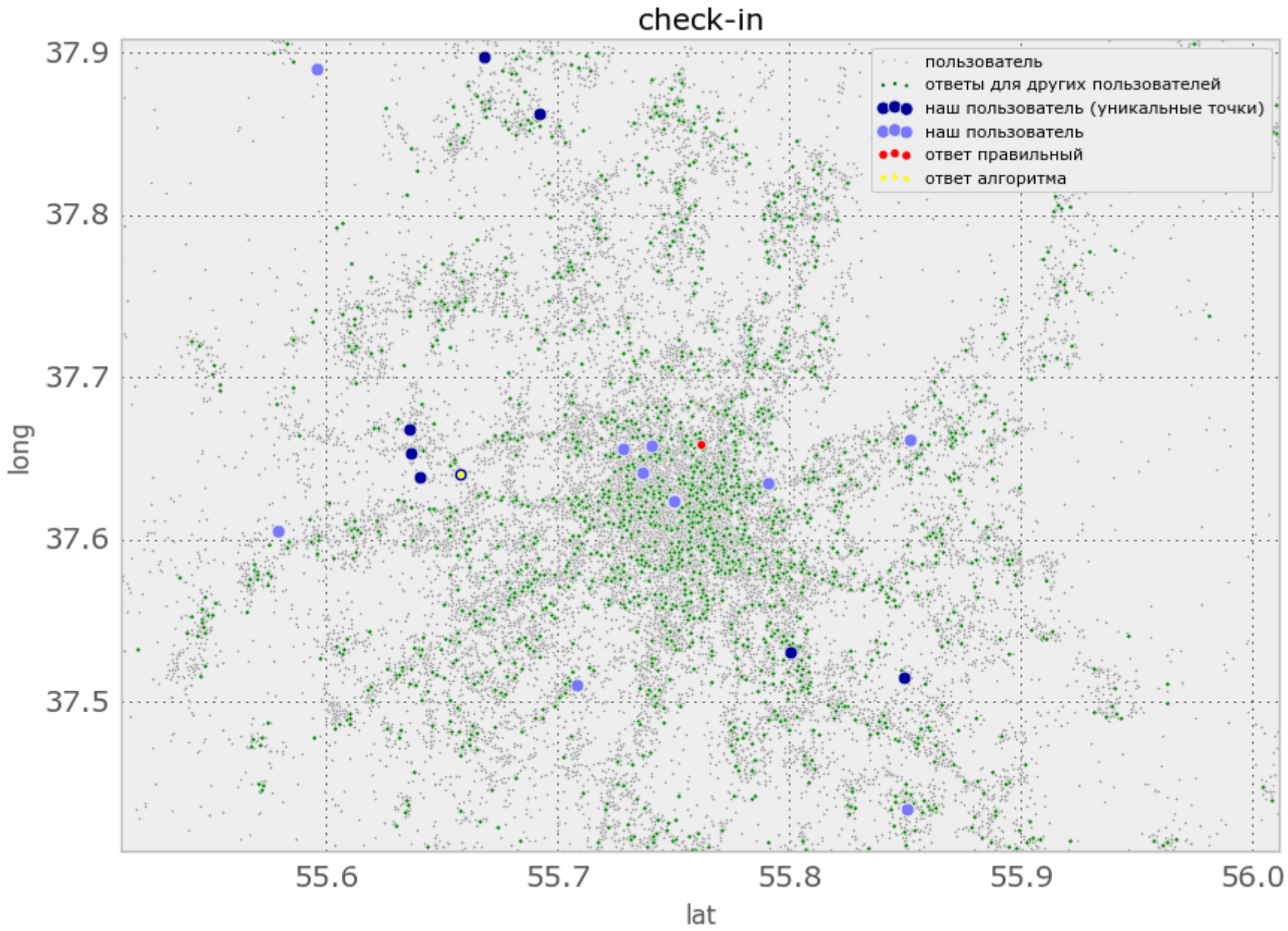
	merchant_id	latitude	longitude	transaction_time	record_time
5824	28477	0.000000	0.000000	2017-01-15 13:02:27	2017-01-15 13:02:20
5825	28477	0.000000	0.000000	2017-01-15 15:44:29	2017-01-15 15:54:15
5826	28477	0.000000	0.000000	2017-01-15 21:33:27	2017-01-15 21:38:17
5827	28477	0.000000	0.000000	2017-01-15 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

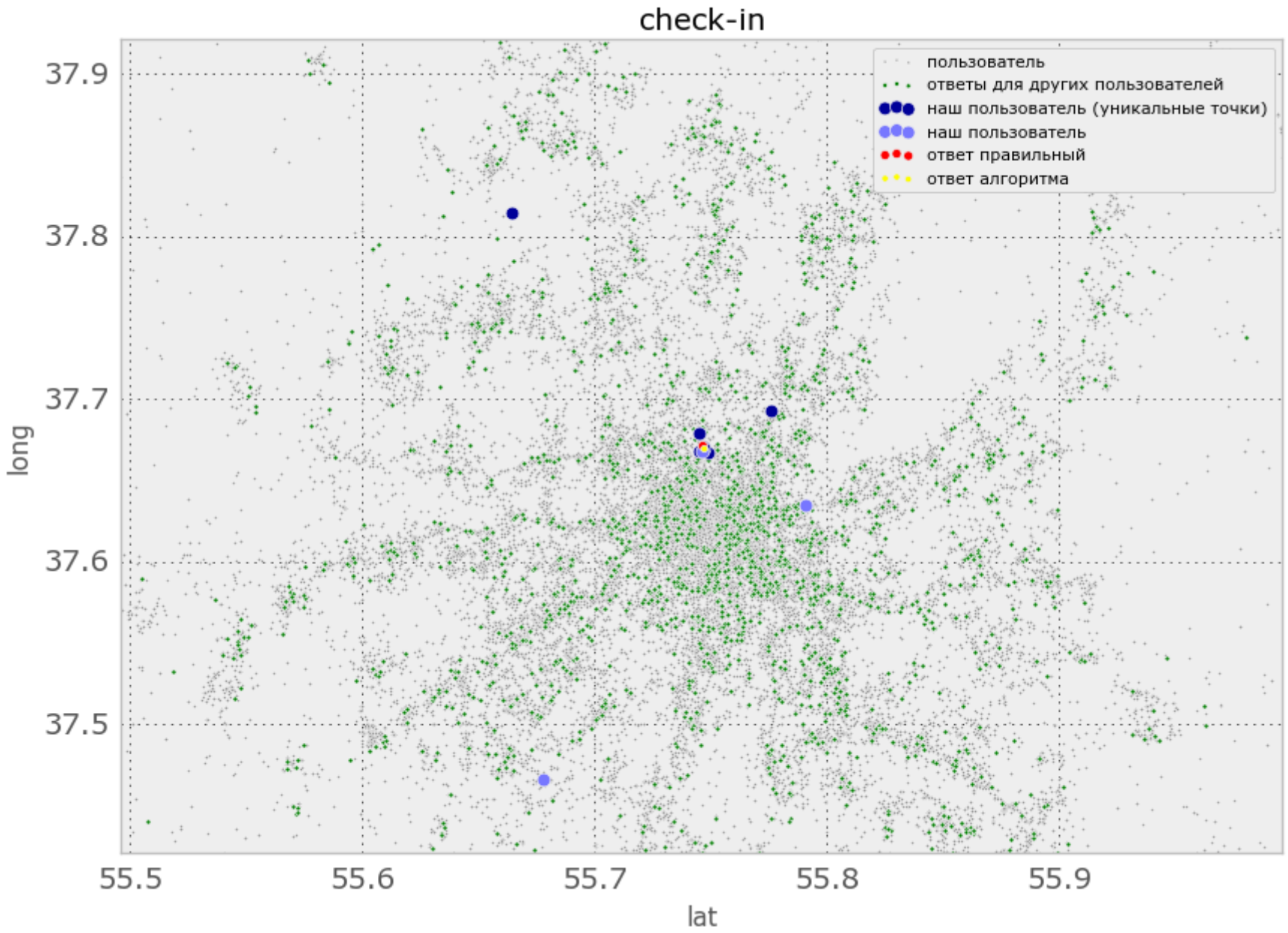
Самый частый check-in

55.75034704 37.62385111 5321




check-in
пользователь
ответы для других пользователей
наш пользователь (уникальные точки)
наш пользователь
ответ правильный
ответ алгоритма
long
37.8
37.7
37.6
37.5
37.4
55.6
55.7
55.8
55.9
56.0
lat


check-in
пользователь
ответы для других пользователей
наш пользователь (уникальные точки)
наш пользователь
ответ правильный
ответ алгоритма
long
37.9
37.8
37.7
37.6
37.5
lat
55.6
55.7
55.8
55.9
56.0


check-in
пользователь
ответы для других пользователей
наш пользователь (уникальные точки)
наш пользователь
ответ правильный
ответ алгоритма
long
37.9
37.8
37.7
37.6
37.5
lat
55.5
55.6
55.7
55.8
55.9

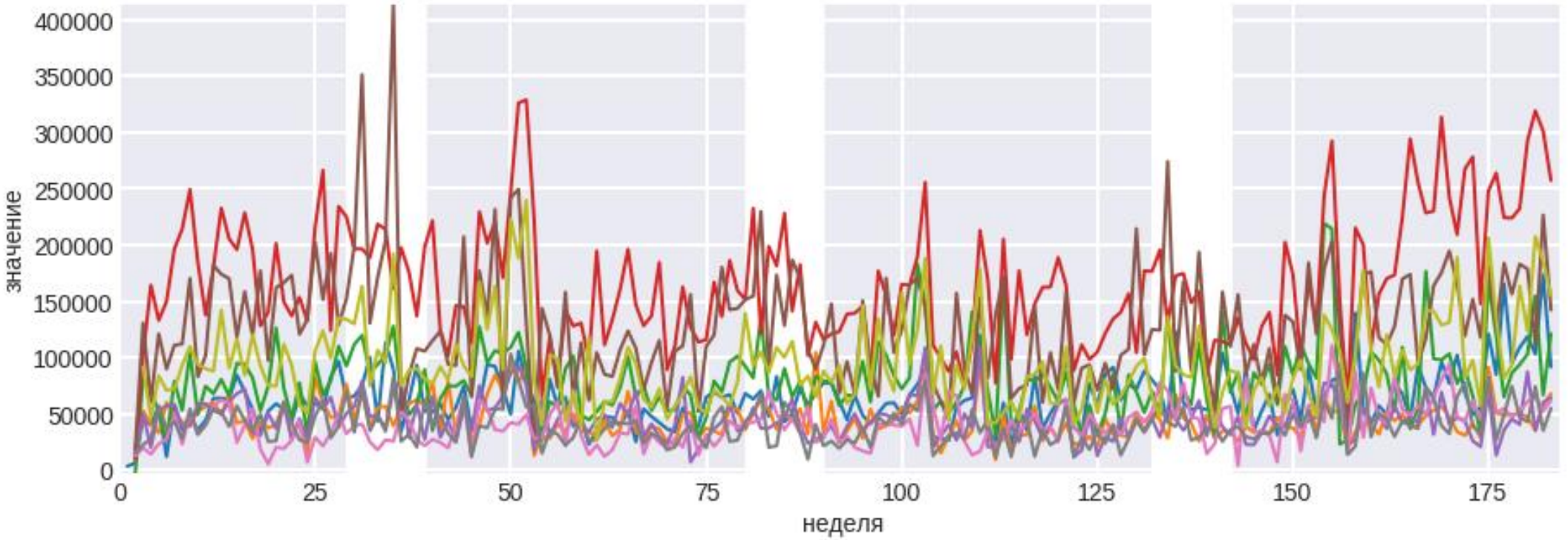
Задача «Train My Data»: планирование продаж Ascott Group

Даны продажи по разным каналам

id	wk	N wk	idFilial	KanalDB	idSubGrp	value
0	201401	1	9	2	3	3560.0
1	201402	2	9	2	3	7120.0
2	201403	3	9	2	3	57672.0
3	201404	4	9	2	3	37380.0
4	201405	5	9	2	3	80990.0
5	201406	6	9	2	3	-8900.0
6	201407	7	9	2	3	131364.0
7	201408	8	9	2	3	67818.0
18247	201722	177	8	1	1	7958.0
18248	201723	178	8	1	1	2076.0
18249	201724	179	8	1	1	8304.0
18250	201725	180	8	1	1	10726.0
18251	201726	181	8	1	1	4152.0
18252	201727	182	8	1	1	2768.0
18253	201728	183	8	1	1	11764.0

Задача

Как выглядят агрегаты по разным каналам…

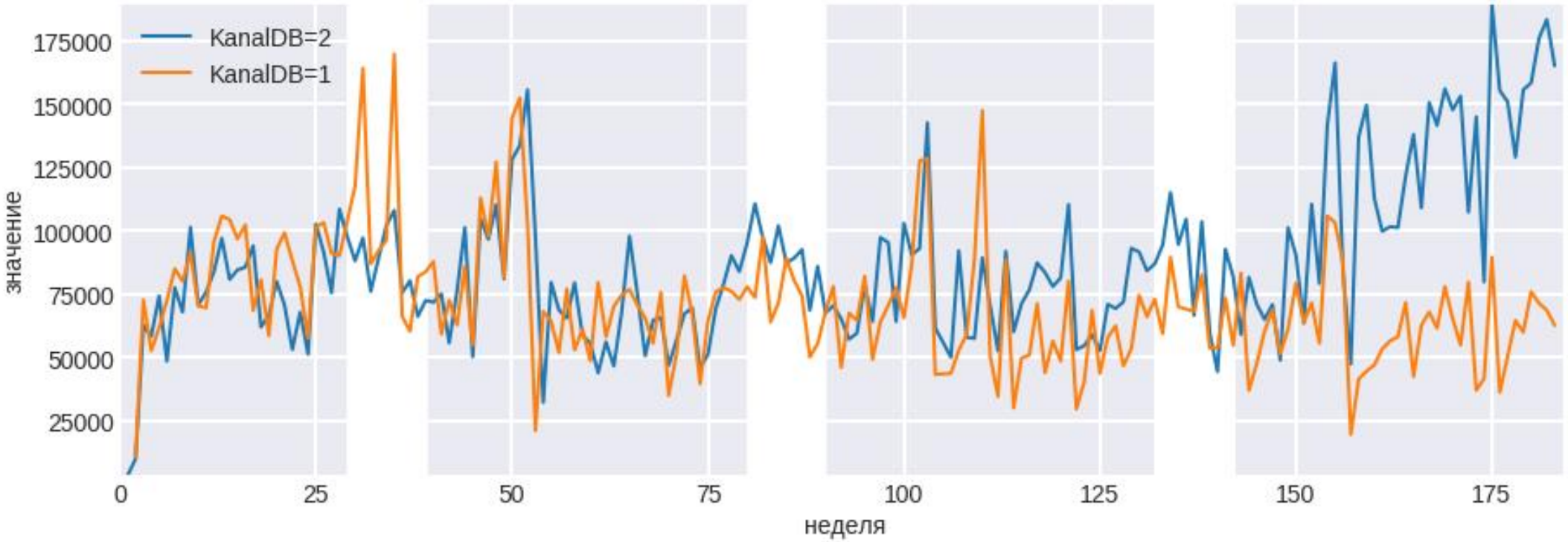


Агрегация по idFilial

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

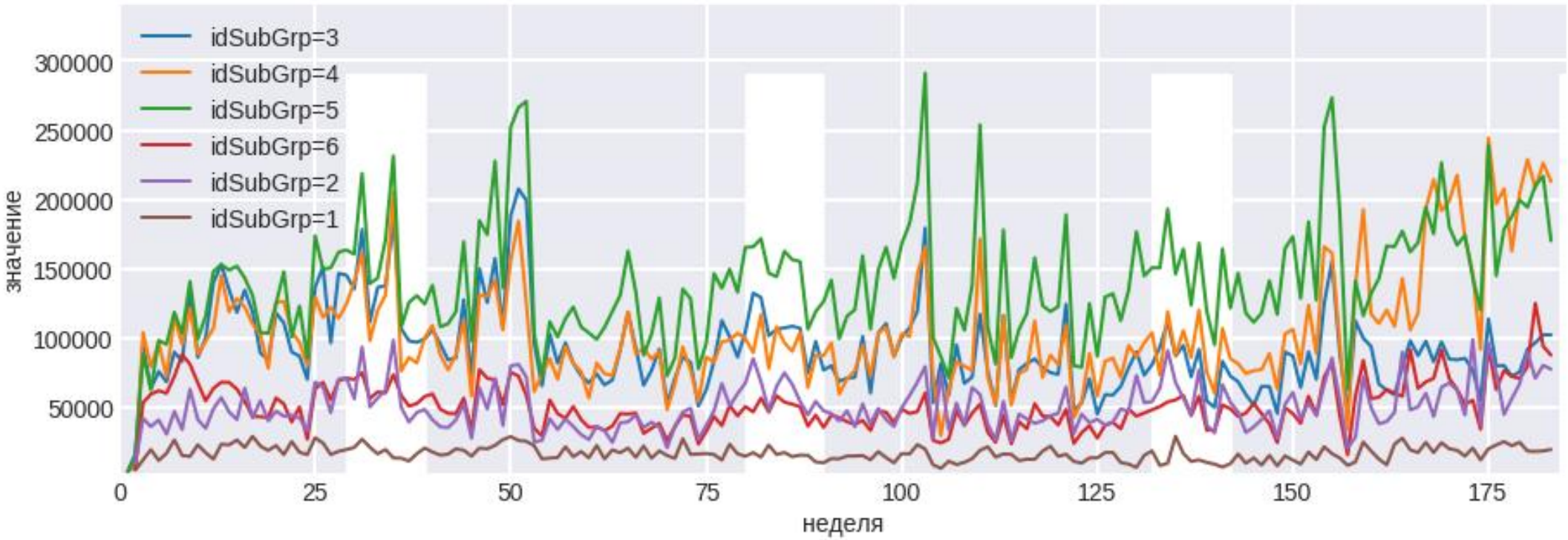
Задача

Как выглядят агрегаты по разным каналам…



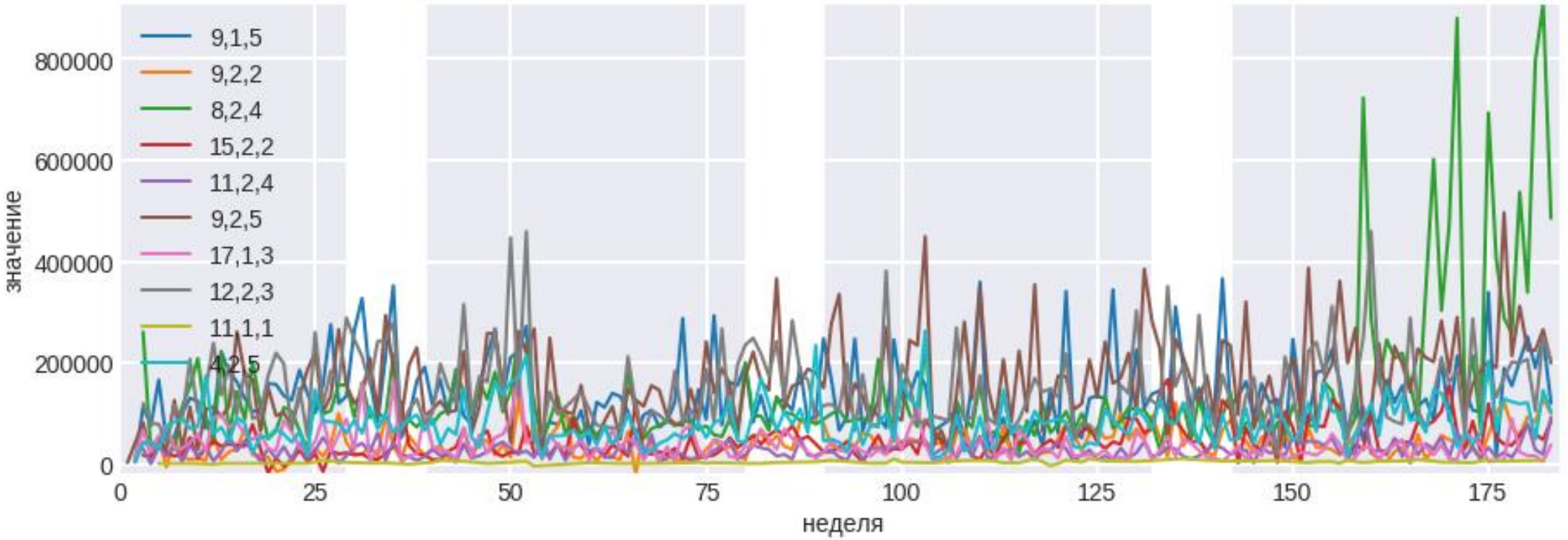
Задача

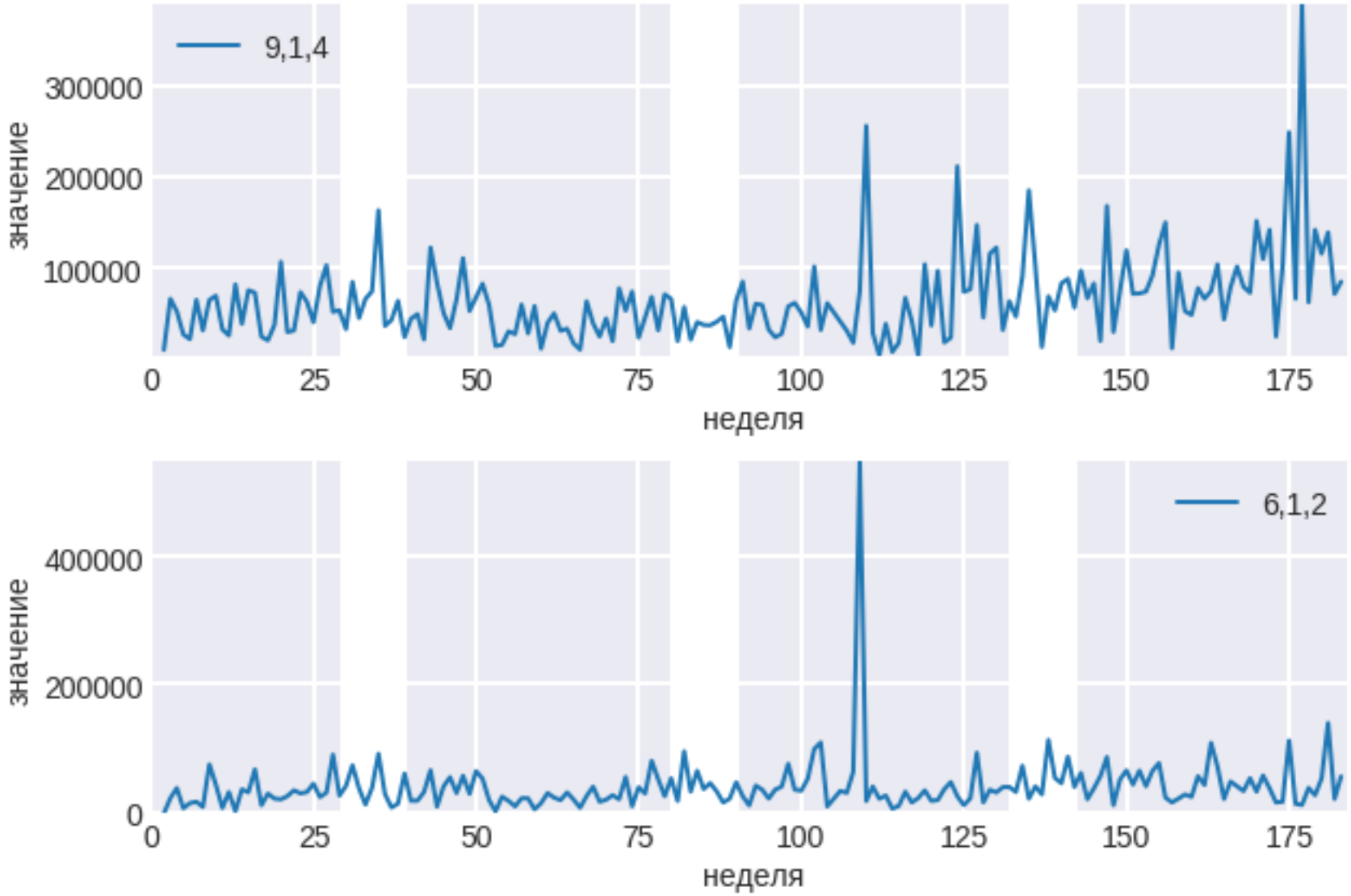
Как выглядят агрегаты по разным каналам…



Задача

Как выглядят прогнозируемые ряды




9,1,4
300000
200000
100000
значение
0
25
50
75
100
125
150
175
неделя
6,1,2
400000
200000
0
значение
0
25
50
75
100
125
150
175
неделя