Методы анализа данных кассовых операций

“Что? Где? Почём? И в каких объёмах?!”

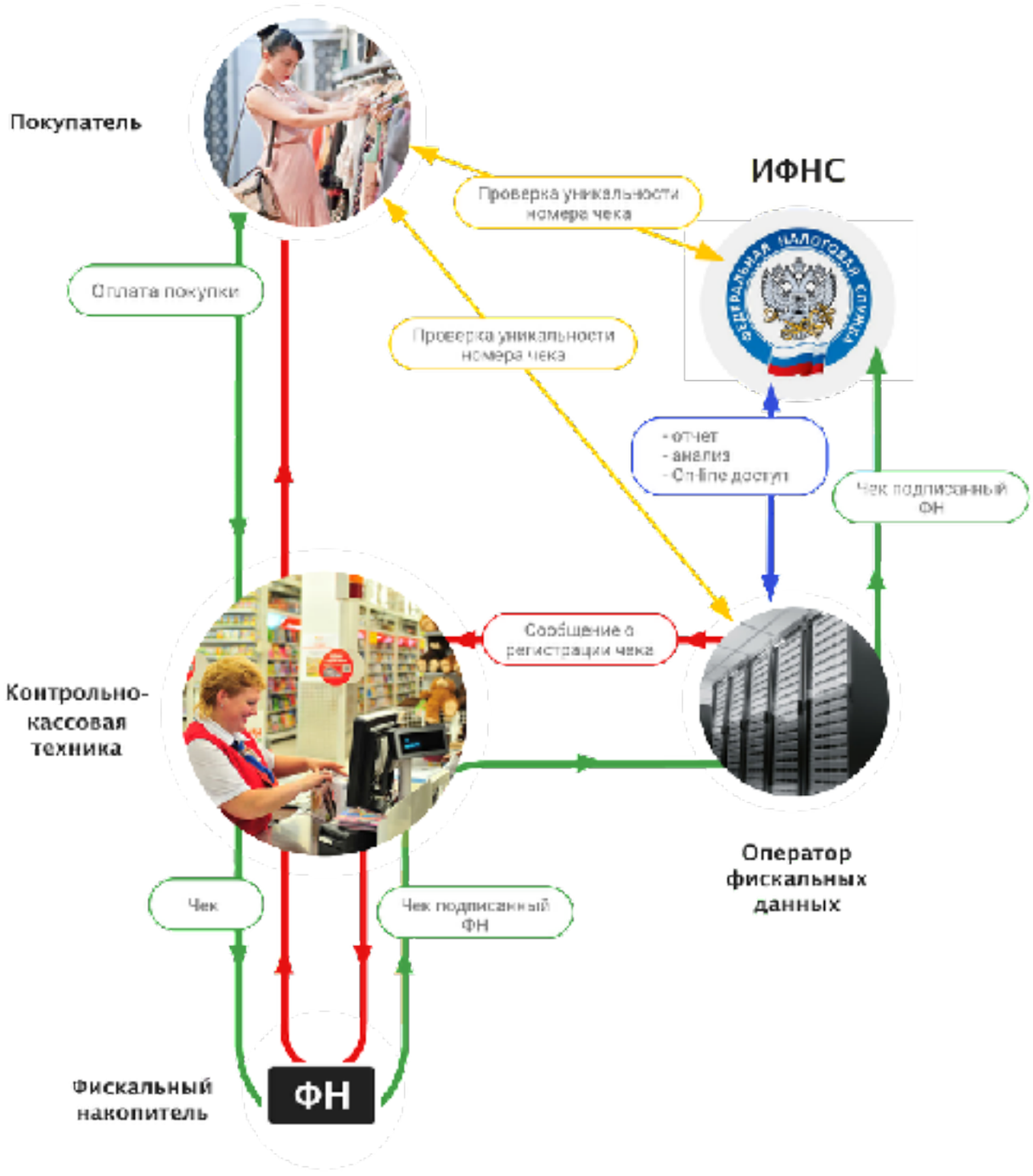
Семинар научно учебной группы
“Анализа мультимидийных данных”

Александр Пономаренко

Нижний Новгород

2018

54-ФЗ: Новый порядок применения ККТ



Источник: http://54-fz.ru

Задачи

- Определение товара по наименованию в чеке
- Автоматическое объединение товарных онтологий
- Поиск мошенничества

Определение товара по наименованию в чеке

$$f : S \to Y$$

S – множество всевозможных строк

Y – множеств товаров

Определение группы товара по наименованию в чеке

```
In [55]: # model.wv.most_similar(positive=['woman', 'king'], negative
         model.wv.most_similar(positive=['мальборро'])

Out[55]: [('марльборо', 0.777965247631073),
          ('"мальборо"', 0.7741876840591431),
          ('marlboro', 0.7476246356964111),
          ('"marlboro"', 0.7189074754714966),
          ('марлборо', 0.6776925325393677),
          ('мальбаро', 0.6735876798629761),
          ('мальбора', 0.6618771553039551),
          ('"marlboro', 0.6340377330780029),
          ('мальбро', 0.6215258836746216),
          ('malboro', 0.6188746094703674)]
```

```
In [407]: model.wv.most_similar(positive=['раптор'])

Out[407]: [('рэйд', 0.6219622492790222),
           ('фумитокс', 0.6086945533752441),
           ('сримоль', 0.6022753119468689),
           ('raid', 0.5635479092597961),
           ('рейд', 0.5542121529579163),
           ('help', 0.5511197447776794),
           ('mosquitall', 0.5343001484870911),
           ('комсрикофо', 0.5232884287834167),
           ('insecto', 0.5229565693305969),
           ('фумигатор-жидкость', 0.5203133225440979)]
```

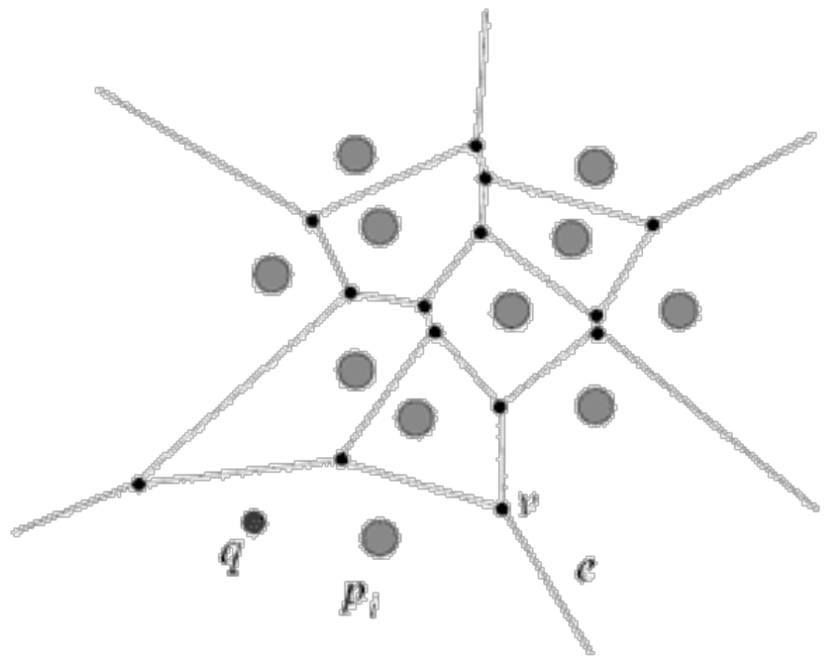
```
In [488]: #модель с окном 5
          model.wv.most_similar(positive=['раптор'])

Out[488]: [('raid', 0.6635477542877197),
           ('3502413', 0.6112322211265564),
           ('сримоль', 0.567592442035675),
           ('рейд', 0.5637607574462891),
           ('3502406', 0.5529232025146484),
           ('off!', 0.5444722771644592),
           ('riad', 0.5408523082733154),
           ('3443803', 0.5371160507202148),
           ('20320', 0.5368900299072266),
           ('3453015', 0.5336128473281860]]
```

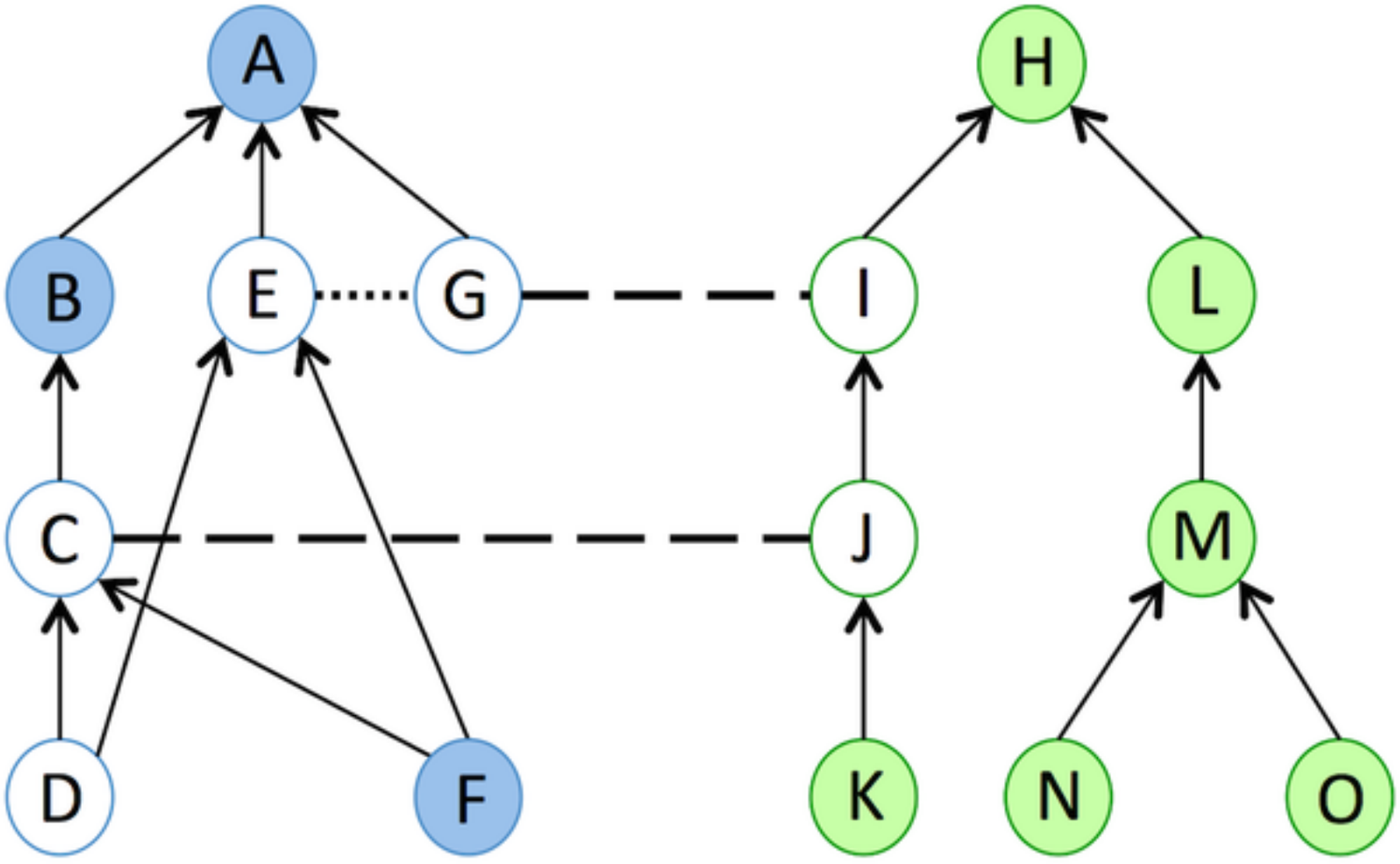
Word2vec на данных 126 миллионах уникальных записей наименований из чеков

Один из хорошо зарекомендовавших подходов для решения задач многоклассовой классификации – это метод k-ближайших соседей.

Одно из главных достоинств этого интерпретируемость.



Сопоставление онтологий

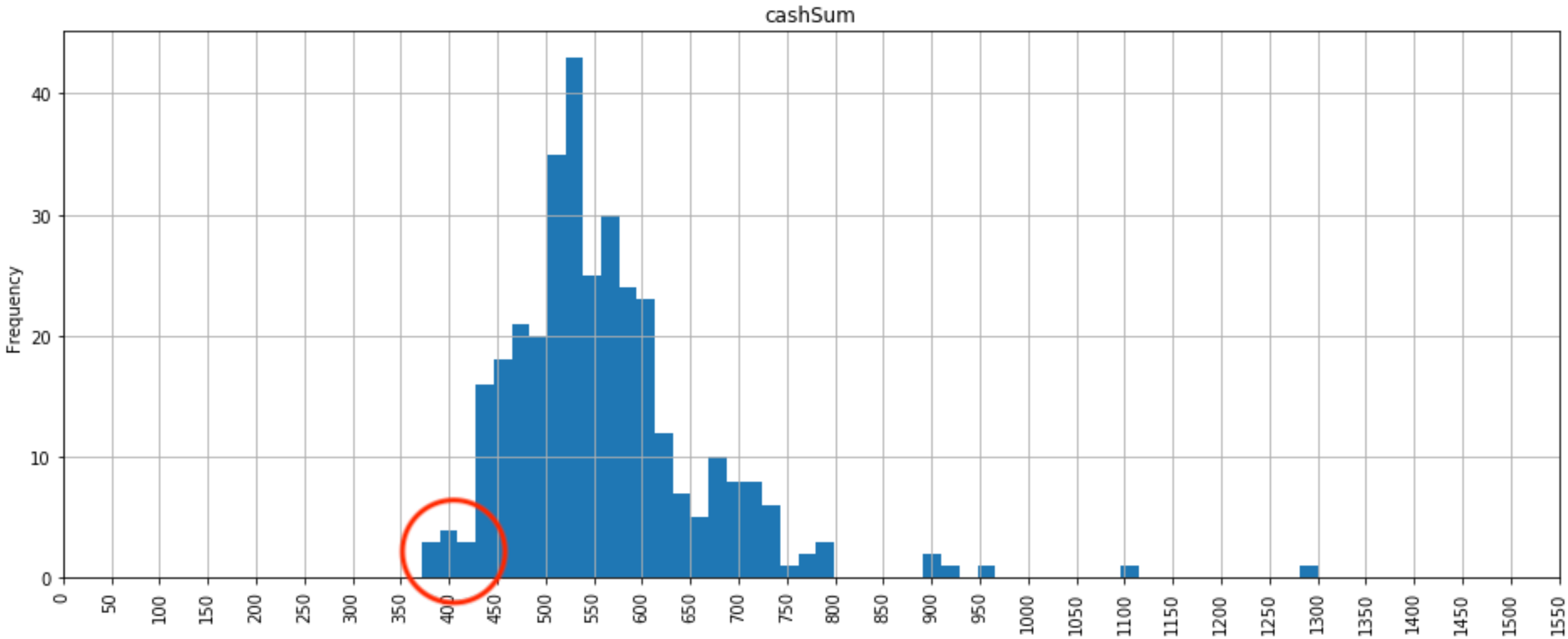


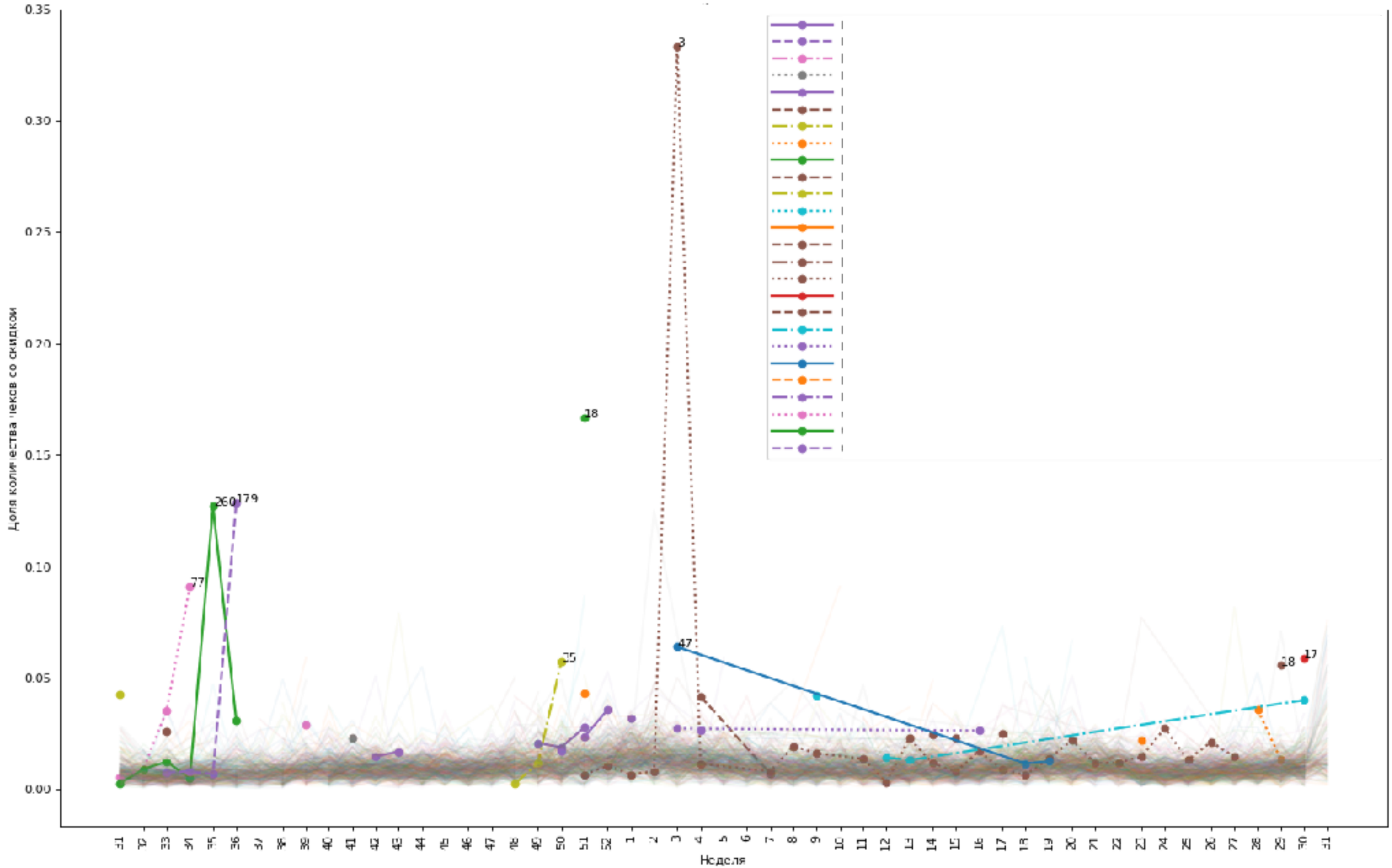
Определение товара по наименованию в чеке

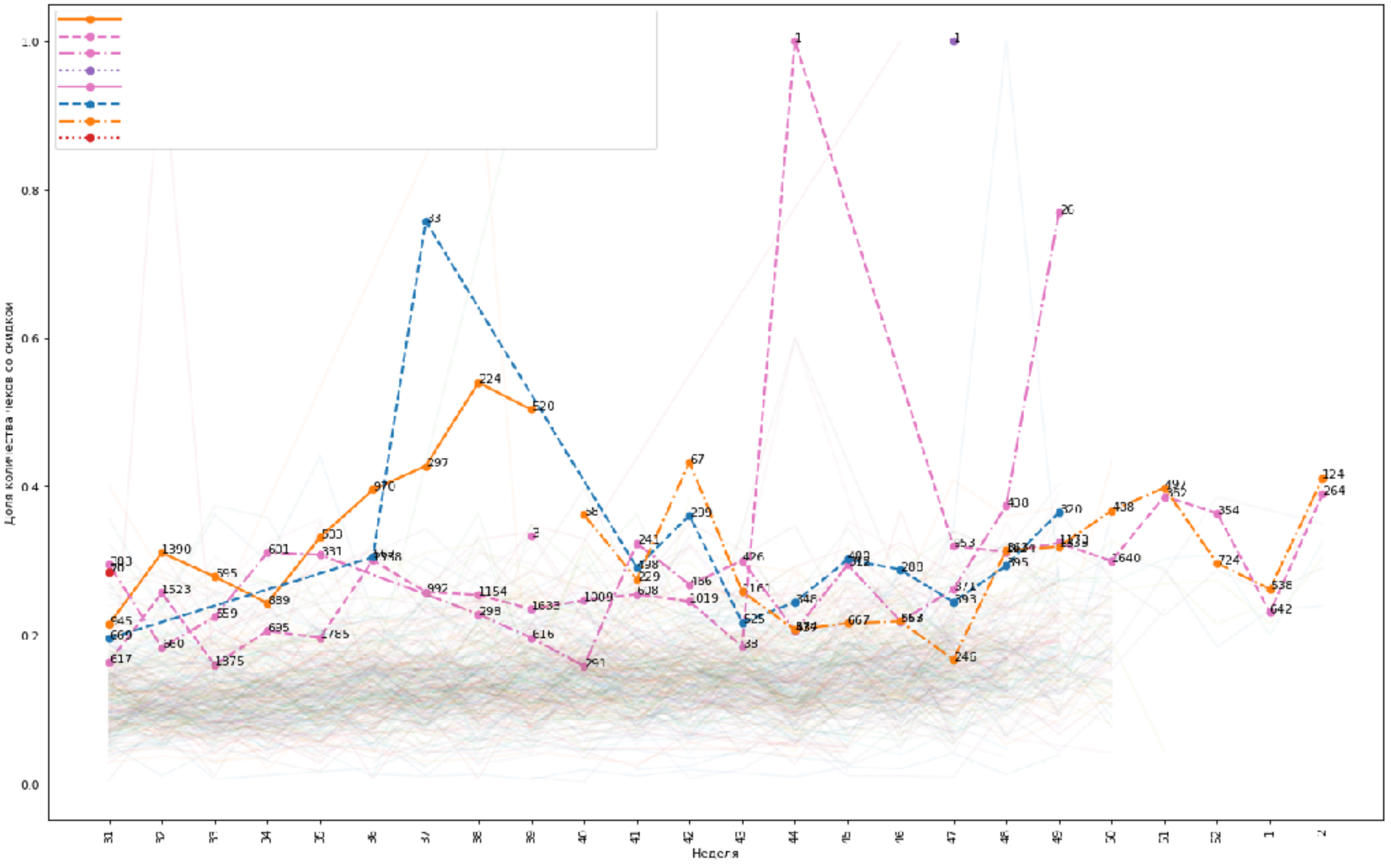
```
…
query : спиралькомарикофф детс
expected: 4607036353023
result: 0.8148148148148148 4607036356215 КОМАРИКОФФ Комплект Детский 45
all:
4: 0.8148148148148148 4607036356215 КОМАРИКОФФ Комплект Детский 45
3: 0.6702508960573477 2113947585535 Полукомбинезон детский арт. AW
2: 0.6543778801843317 2110001724220 Полукомбинезон дет. арт.AW12-L
1: 0.6527777777777778 4600694114012 Подарок детский новогодний Чем
0: 0.6444444444444444 4603290000031 Печенье детское Чипа Экстрамал
--------------------------------------------------
query : 049762 раптор спираль от комар.б/з10ш
expected: 8025051349054
result: 0.7999999999999999 4607131422327 Раптор спираль от комаров стандарт 10шт
all:
4: 0.7999999999999999 4607131422327 Раптор спираль от комаров стандарт 10шт
3: 0.7037037037037037 4607131422365 РАПТОР Пластины от комаров 10ш
2: 0.7000000000000001 4620000430353 RAID Спираль от комаров 10шт Россия
1: 0.6787878787878787 4607036353023 комарикофф Спирали от комаров малые 10шт детские
0: 0.6666666666666666 4607131422310 РАПТОР ПРОМО Спираль от комаро
--------------------------------------------------
query : спираль рейд 5*2 шт 24*10
expected: 5010182988442
result: 0.5833333333333333 4605865820143 Набор полотенец махр. 2 шт 50х
all:
4: 0.5833333333333333 4605865820143 Набор полотенец махр. 2 шт 50х
3: 0.5702630702630702 2190414280520 Гирлянда из фольги 12.5 см х 2
2: 0.5445887445887445 2113730933963 Пижама дет. д/дев. арт.104 р.2
1: 0.5357142857142857 4680252015838 КПБ 2.0 сп ДАЧНЫЙ. нав 50*70 и
0: 0.5336734693877551 4810096219004 Клатч ИК 245*10*130мм арт. 314

Точность: 0.969388
CPU times: user 2h 20min 2s, sys: 4.66 s, total: 2h 20min 6s
Wall time: 2h 20min 7s
```

Выявление мошенничества




0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Доля количества чеков со скидкой
Неделя
3
18
260
179
77
25
47
18
17


1.0
0.8
0.6
0.4
0.2
0.0
Доля количества чеков со скидкой
Неделя
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
1
2