Итоговая аттестация: Проект 9. «Сегментация клиентской базы и поведенческое профилирование»

Богатырева Ксения, Матвеев Максим

Рабочая группа

Ксения Богатырева

Задачи:
1) Анализ и чистка данных
2) Кластеризация методом K-Means
3) Оценка качества кластеризации
4) Подготовка презентации

Максим Матвеев

Задачи:
1) Кластеризация EM алгоритмом
2) Оценка качества кластеризации
3) Подготовка презентации

Цель: Определить целевые сегменты на основе моделей кластеризации.

Задачи :

- Исследование качества данных;
- Подготовка витрин для моделей сегментации;
- Построение моделей сегментации;
- Определение оптимального кол-во сегментов на основе статистических методов и проверка построенной модели сегментации на качество;
- Описание полученных сегментов по каждой модели - выделение профилей по клиентам.

Описание данных. Введение

Представлены:

Данные о размещенных и выкупленных заказах в интернет магазине бытовых товаров.

Состав данных:

37 переменных:

- 14 количественных
- 22 категориальных

357036 записей

Order_ID	Email_new	Phone_new	Source	OrderDate	время	месяц	ChangeDate	DeliveryDate	...
1303000509_TT	55666668105117_iu29@yandex.ru	55485656-57565656575275	Онлайн-Резерв.	2016-03-30	08:46:30.000	201603	2016-03-30 09:31:57.000	2016-04-06 00:00:00.000	
1303000509_TT	55666668105117_iu29@yandex.ru	55485656-57565656575275	Онлайн-Резерв.	2016-03-30	08:46:30.000	201603	2016-03-30 09:31:57.000	2016-04-06 00:00:00.000	
1303000509_TT	55666668105117_iu29@yandex.ru	55485656-57565656575275	Онлайн-Резерв.	2016-03-30	08:46:30.000	201603	2016-03-30 09:31:57.000	2016-04-06 00:00:00.000	

Анализ данных

Кол-во уникальных значений

Кол-во нулевых значений

Кол-во пустых значений

Column	Count Unique	Count Zeros	Count NaNs	% of NaNs	data type
Order_ID	166794	0	0	0	object
Email_new	99284	61511	0	0	object
Phone_new	123135	7565	0	0	object
Source	4	0	0	0	object
OrderDate	61	0	0	0	datetime64[ns]
Время	58549	0	0	0	object
Месяц	2	0	0	0	int64
ChangeDate	155189	0	0	0	object
DeliveryDate	100	0	0	0	object
PaymentDate	105118	0	0	0	object
Status	15	0	0	0	object
Status_ID	15	0	0	0	int64
OneClick	2	326049	0	0	int64
CancelReason	34	0	230861	64.7	object
Actions	41	0	261490	73.2	object
DeliveryType	2	0	8285	2.3	object
PaymentType	7	0	0	0	object
Region	73	0	2	0	object
Area	61	0	1631	0.5	object
Store_ID	167	965	0	0	int64
FullSum	15995	65	10	0	float64
Discount	1920	309994	10	0	float64
IM_Rozn_Sum	20622	5574	10	0	float64
Row_ID	41	0	10	0	float64
Articul	26930	0	10	0	float64
Nom_Name	26888	0	48	0	object
NomGroup	849	0	10	0	object
Quant	43	5	10	0	float64
RowPrice	9601	152336	10	0	float64
RowDiscount	2006	333347	10	0	float64
RowSum	12377	152813	10	0	float64
Brand	1179	0	166131	46.5	object
TN	17	0	166059	46.5	object
TK	112	0	166059	46.5	object
NomFullPath	878	0	10	0	object
Week	9	0	10	0	float64
Nom_ID	26928	3	10	0	float64

2 идентификатора клиента. “Email_new” имеет больше пропущенных знач.

2 месяца: 03-04.2016

Акционных товаров куплено намного меньше

2 похожих показателя локации

Товаров по скидке очевидно куплено намного меньше

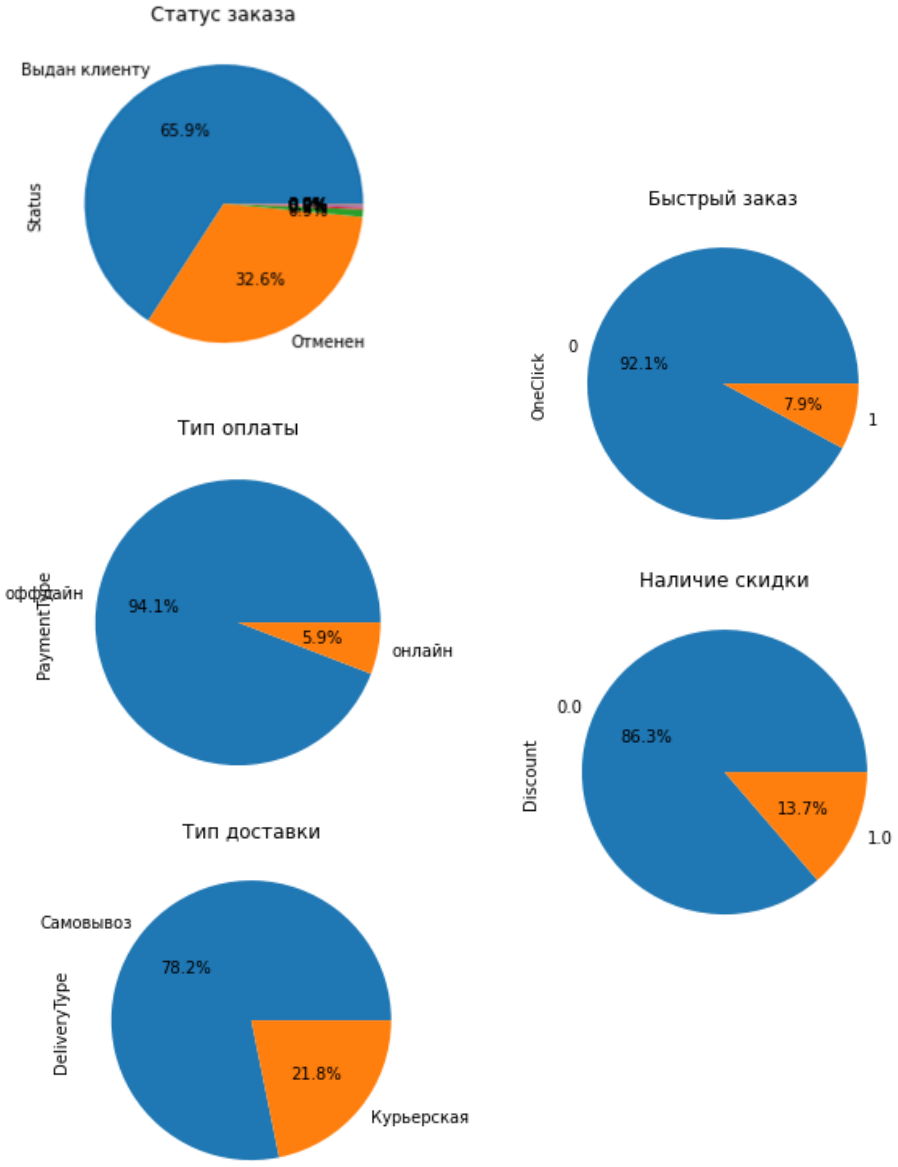
TN, TK – много пустых значений, которые можно заполнить с помощью NomFullPath

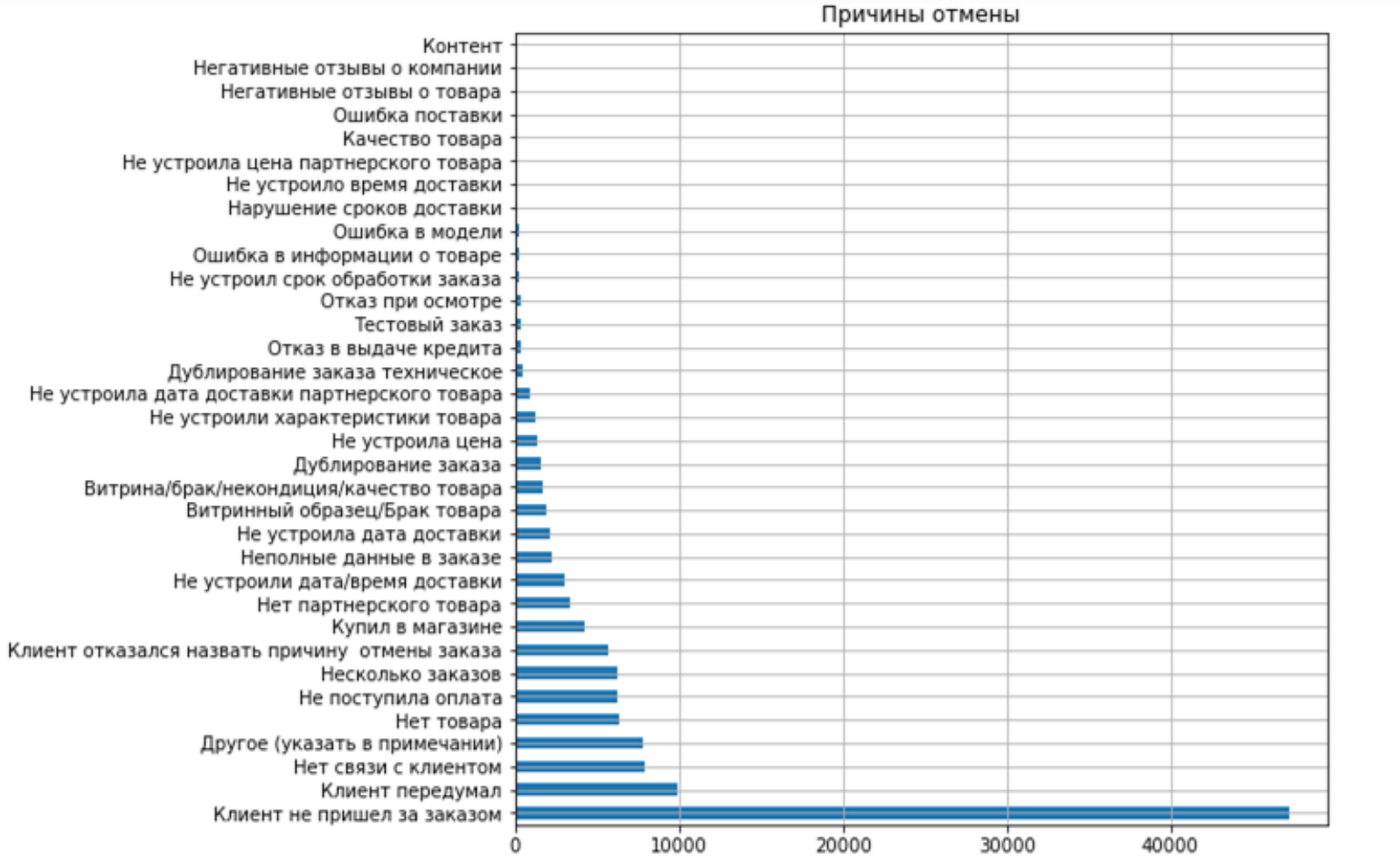
Тип “Object” преобладает

Быстрых заказов намного меньше

Отмененных заказов меньше

Анализ данных





Очистка данных

Тип	Параметр	Расшифровка параметра	Изменения
Кол	Order_ID	Номер чека	
Кач	Email_new	Эл. адрес клиента	Удален, так как идентификатором клиента выбран параметр «Phone_new»
Кач	Phone_new	Моб. Тел. клиента	Удалены 7565 (2%) записей с пропущенными значениями
Кол	время	Время создания заказа	
Кол	месяц	Месяц создания заказа	
Кол	ChangeDate	Дата внесения изменений	
Кол	DeliveryDate	дата доставки	
Кол	PaymentDate	дата оплаты	
Кач	Status	статус заказа	
Кач	Status_ID	id_статуса заказа	
Кач	OneClick	флаг совершения быстрого заказа	
Кач	CancelReason	причина отмены	Удалены 8588 записей (2%): значения «Дублирование заказа», «Дублирование заказа техническое», «Тестовый заказ», «Несколько заказов»
Кач	Actions	акции	
Кач	DeliveryType	тип доставки	
Кач	PaymentType	тип оплаты	Значения были заменены на «онлайн» и «оффлайн», где «онлайн» - моментальная онлайн оплата в момент создания заказа (5,9%), «оффлайн» - оплата в момент забора заказа (94,1%)
Кач	Region	Город	Заменен на более широкий параметр «Регион»
Кач	Area	Область	Заменен на более широкий параметр «Регион»
Кач	Store_ID	ID_склада	
Кол	FullSum	Сумма чека с учетом всех скидок	

Очистка данных

Тип	Параметр	Расшифровка параметра	Изменения
Кол	Discount	Общая сумма скидки по чеку	Значения были заменены на «0» и «1», где «0» - без скидки (86,3%), «1» - со скидкой (13,7%)
Кач	NomFullPath	Детализированное название группы товаров	Удалены 158333 записей (44%): доставка, установка и настройка техники, страхование и пр. услуги
Кол	IM_Rozn_Sum	Общая сумма чека без учета скидок	
Кач	Row_ID	Поряд. номер товара в чеке	
Кач	Articul	Номер артикула	
Кач	Nom_Name	Детализированное название товара	
Кач	NomGroup	Обобщенное название товара	Удалены 40 записей (0%): значения «Номенклатура к привязке»
Кол	Quant	Количество товаров	
Кол	RowPrice	Цена товара без скидки	
Кол	RowDiscount	Сумма скидки на товар в чеке с учетом его кол-ва	
Кол	RowSum	Выручка	
Кач	Brand	Бренд	
Кач	TN	Товарное направление	
Кач	TK	Название группы товаров	Удалены 12 записей (0 %): значение «Подарки КБТ»
Кол	Week	Неделя создания заказа	
Кач	Nom_ID	ID номенклатуры	
Кач	Регион	Регион	Параметр был присоединен из другого датасета

Итог: Датасет после чистки: 182 479 записей (удалено 174 557 записей(49%))

Сводные таблицы

Группировка по товарным направлениям

Выданные заказы

TN	Выручка	Кол-во чеков	Кол-во уникальных клиентов	Средний чек	Среднее кол-во товаров в чеке	Кол-во товаров	Кол-во уникальных товаров
Мелкая бытовая техника	128820535	24574	21844	6347	1.023184	26392	76
Крупная бытовая техника	357532141	17859	14795	23303	1.008536	19613	27
Цифровая Техника	148919945	11694	10923	13660	1.033031	12416	40
Сделай сам	67222249	9908	9179	7704	1.051888	12650	155
Аксессуары	18020355	9795	9010	5051	1.081550	11724	138
Спорт и активный отдых	58498313	8735	8112	7208	1.131550	11509	79
Детские товары	44753951	7464	7095	6277	1.022604	8460	139
ТВ-Аудио	135011848	6923	5825	20749	1.016695	7186	38
Мебель	24643637	5272	5015	5674	1.110926	6670	38
Товары для дома	8435561	4479	4085	4132	1.152154	6391	125
Климат	18310733	2255	2155	10133	1.093258	2708	14
Игры, софт и развлечения	1263014	370	352	4544	1.012225	414	18
Услуги	150415	78	62	2241	2.011905	169	5
Автотовары	83000	50	49	1984	1.020000	51	2
Элитная техника	139878	3	3	46626	1.000000	3	2
Хобби, досуг	35	1	1	2837	1.000000	1	1

Размещенные заказы

TN	Выручка	Кол-во чеков	Кол-во уникальных клиентов	Средний чек	Среднее кол-во товаров в чеке	Кол-во товаров	Кол-во уникальных товаров
Мелкая бытовая техника	200772550	34925	28881	7875	1.046670	39247	80
Крупная бытовая техника	554045314	26250	19647	24784	1.012159	29053	27
Цифровая Техника	367147113	21280	18170	18715	1.040267	22941	42
Сделай сам	108408957	14440	12446	8766	1.059768	18352	158
Аксессуары	28201988	14020	12097	6315	1.102765	17427	144
Спорт и активный отдых	100979470	13597	11331	8297	1.172427	18760	85
Детские товары	75637585	11690	10054	6914	1.034803	13380	143
ТВ-Аудио	247711206	11184	8690	24179	1.103005	12775	39
Мебель	34128969	6933	6063	6222	1.111827	8799	40
Товары для дома	12554811	6310	5324	4610	1.173519	9191	129
Климат	31055338	3268	2870	13840	1.117942	4237	15
Игры, софт и развлечения	4654384	655	600	12508	1.058065	820	21
Услуги	371952	169	125	2740	1.791209	326	5
Автотовары	171846	79	71	3236	1.012658	80	3
Элитная техника	304308	6	6	50718	1.000000	6	4
Хобби, досуг	35	1	1	2837	1.000000	1	1
Инсталляция	6270	1	1	6270	30.000000	30	1

Выкупаемость

TN	Выручка	Кол-во чеков	Кол-во уникальных клиентов	Средний чек	Среднее кол-во товаров в чеке	Кол-во товаров	Кол-во уникальных товаров
Хобби, досуг	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Мебель	72.21%	76.04%	82.71%	91.19%	99.92%	75.80%	95.00%
Товары для дома	67.19%	70.98%	76.73%	89.63%	98.18%	69.54%	96.90%
Мелкая бытовая техника	64.16%	70.36%	75.63%	80.60%	97.76%	67.25%	95.00%
Аксессуары	63.90%	69.86%	74.48%	79.99%	98.08%	67.27%	95.83%
Климат	58.96%	69.00%	75.09%	73.22%	97.79%	63.91%	93.33%
Сделай сам	62.01%	68.61%	73.75%	87.89%	99.26%	68.93%	98.10%
Крупная бытовая техника	64.53%	68.03%	75.30%	94.03%	99.64%	67.51%	100.00%
Спорт и активный отдых	57.93%	64.24%	71.59%	86.88%	96.51%	61.35%	92.94%
Детские товары	59.17%	63.85%	70.57%	90.79%	98.82%	63.23%	97.20%
Автотовары	48.30%	63.29%	69.01%	61.33%	100.72%	63.75%	66.67%
ТВ-Аудио	54.50%	61.90%	67.03%	85.82%	92.18%	56.25%	97.44%
Игры, софт и развлечения	27.14%	56.49%	58.67%	36.33%	95.67%	50.49%	85.71%
Цифровая Техника	40.56%	54.95%	60.12%	72.99%	99.30%	54.12%	95.24%
Элитная техника	45.97%	50.00%	50.00%	91.93%	100.00%	50.00%	50.00%
Услуги	40.44%	46.15%	49.60%	81.77%	112.32%	51.84%	100.00%
Инсталляция	nan%	nan%	nan%	nan%	nan%	nan%	nan%

Сводные таблицы

Группировка по регионам

Размещенные заказы

Регион	Выручка	Кол-во чеков	Кол-во уникальных клиентов	Средний чек	Среднее кол-во товаров в чеке	Кол-во товаров	Кол-во уникальных товаров
CENTRAL	930418763	83373	65481	12535	1.077961	103246	792
NORTH	415442181	39165	29143	12301	1.045838	46362	748
PRIVOLZIE	207582040	18039	13732	14002	1.069605	22374	688
SIBERIA	83732472	7795	6337	12054	1.094007	9729	581
SOUTHERN	63993905	5542	4370	12537	1.102253	7093	545
URAL	64016803	5195	3917	15489	1.083265	6583	535
FAR EAST	1027500	42	42	21902	1.089286	61	42

Выданные заказы

Регион	Выручка	Кол-во чеков	Кол-во уникальных клиентов	Средний чек	Среднее кол-во товаров в чеке	Кол-во товаров	Кол-во уникальных товаров
CENTRAL	531390255	55809	47906	10541	1.053350	67644	766
NORTH	252639347	26855	22005	10729	1.039261	31394	692
PRIVOLZIE	113890000	11366	9713	10788	1.043717	13322	621
SIBERIA	45971740	4835	4316	10403	1.077751	5905	514
SOUTHERN	35441134	3467	3015	10885	1.082147	4334	474
URAL	31953242	3032	2538	12787	1.061143	3714	468
FAR EAST	537722	33	33	15485	1.106383	52	35

Выкупаемость

Регион	Выручка	Кол-во чеков	Кол-во уникальных клиентов	Средний чек	Среднее кол-во товаров в чеке	Кол-во товаров	Кол-во уникальных товаров
FAR EAST	52.33%	78.57%	78.57%	70.70%	101.57%	85.25%	83.33%
NORTH	60.81%	68.57%	75.51%	87.22%	99.37%	67.71%	92.51%
CENTRAL	57.11%	66.94%	73.16%	84.09%	97.72%	65.52%	96.72%
PRIVOLZIE	54.87%	63.01%	70.73%	77.05%	97.58%	59.54%	90.26%
SOUTHERN	55.38%	62.56%	68.99%	86.82%	98.18%	61.10%	86.97%
SIBERIA	54.90%	62.03%	68.11%	86.31%	98.51%	60.69%	88.47%
URAL	49.91%	58.36%	64.79%	82.56%	97.96%	56.42%	87.48%

Создание витрины по клиентам

85522 записей × 39 параметров

Финансовые показатели по размещенным и выкупленным заказам

- Выручка
- Кол-во чеков
- Средний чек
- Среднее кол-во товаров в чеке
- Кол-во товаров
- Доля выкупленных

Дополнительные показатели

- Доля самовывоза
- Доля онлайн оплаты
- Доля быстрого заказа
- Геолокация заказов

Статистика по тов. направлениям

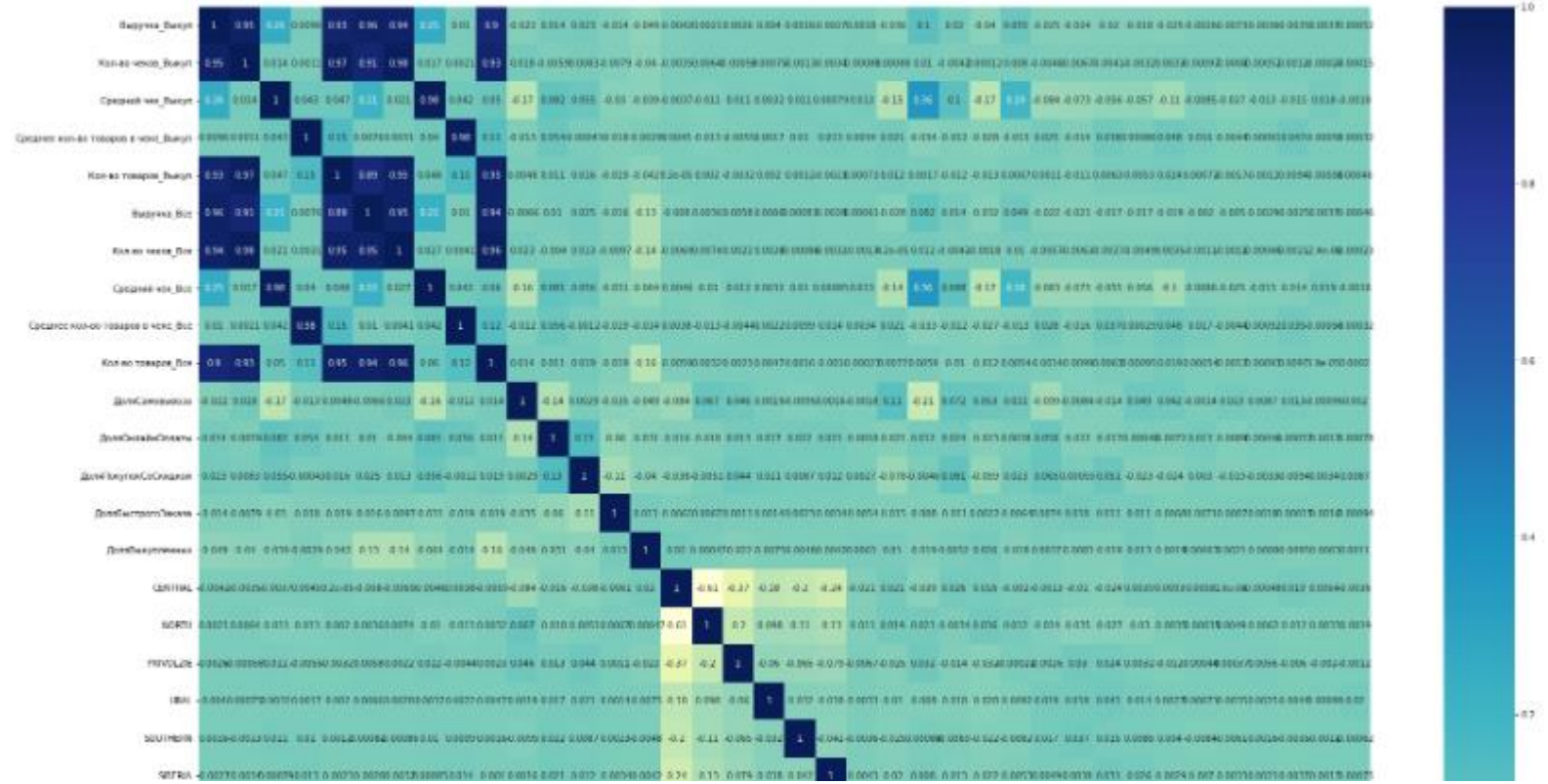
- Доля выкупленных заказов по каждому товарному направлению

Id	Выручка_Выкуп	Кол-во чеков_Выкуп	Средний чек_Выкуп	Среднее кол-во товаров в чеке_Выкуп	Кол-во товаров_Выкуп	...	Автотовары	Услуги	Элитная техника	Хобби, досуг
5575449-54535553535073	2420.0	1	2420.0	1.0	1.0	...	0.0	0.0	0.0	0.0
5574954-53565052504871	8999.0	1	8999.0	1.0	1.0	...	0.0	0.0	0.0	0.0
5575049-51505248534972	1790.0	1	1790.0	1.0	1.0	...	0.0	0.0	0.0	0.0
5574954-53495654564877	1420.0	1	1420.0	1.0	1.0	...	0.0	0.0	0.0	0.0
5575054-53544849485671	249.0	1	249.0	1.0	1.0	...	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...

Выявление значимых параметров

Корреляционная матрица:
выявления статистически значимых связей на основании значения критерия Пирсона

```
fig = plt.figure(figsize = (30, 30))
sns.heatmap(vitrinaTN.corr(), annot=True, cmap = 'YlGnBu');
```



Интерпретация значений:

r_{xy} < 0,3 - *слабая* связь

r_{xy} от 0,3 до 0,85 - связь *средней* тесноты

r_{xy} > 0,85 - *сильная* связь – необходимо удалить

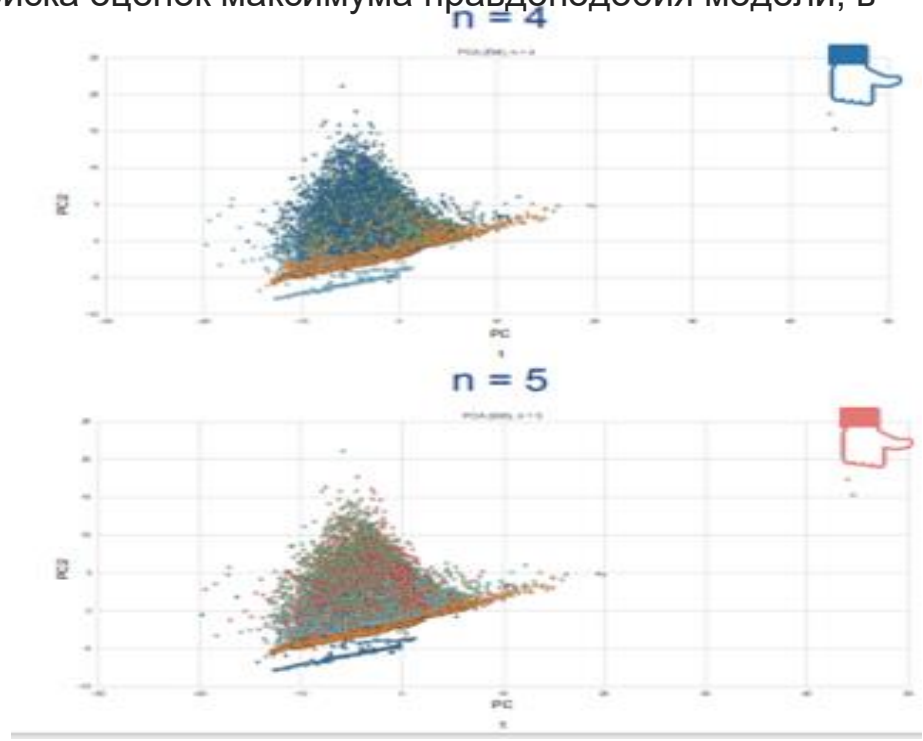
Удаленные параметры:

'Средний чек_Выкуп',
'Выручка_Все',
'Средний чек_Все',
'Среднее кол-во товаров в чеке_Все',
'Кол-во товаров_Все',
'Кол-во чеков_Все'

Метод кластеризации: EM - algorithm

Алгоритм EM (англ. *expectation-maximization*) — итеративный алгоритм поиска оценок максимума правдоподобия модели, в ситуации, когда она зависит от скрытых (ненаблюдаемых) переменных.

1. Алгоритм ищет параметры модели итеративно, каждая итерация состоит из двух шагов:
2. **E (Expectation)** шаг — поиск наиболее вероятных значений скрытых переменных.
3. **M (Maximization)** шаг — поиск наиболее вероятных значений параметров, для полученных на шаге E значений скрытых переменных.
4. EM алгоритм подходит для решения задач двух типов:
5. 1.Задачи с неполными данными.
   2.Задачи, в которых удобно вводить скрытые переменные для упрощения подсчета функции правдоподобия. Примером такой задачи может служить кластеризация.



+

- Сходится в большинтсве случаев.
- Наиболее гибкое решение.
- Существуют простые модификации, позволяющие уменьшить чуствительность алгоритма к шуму в данных.

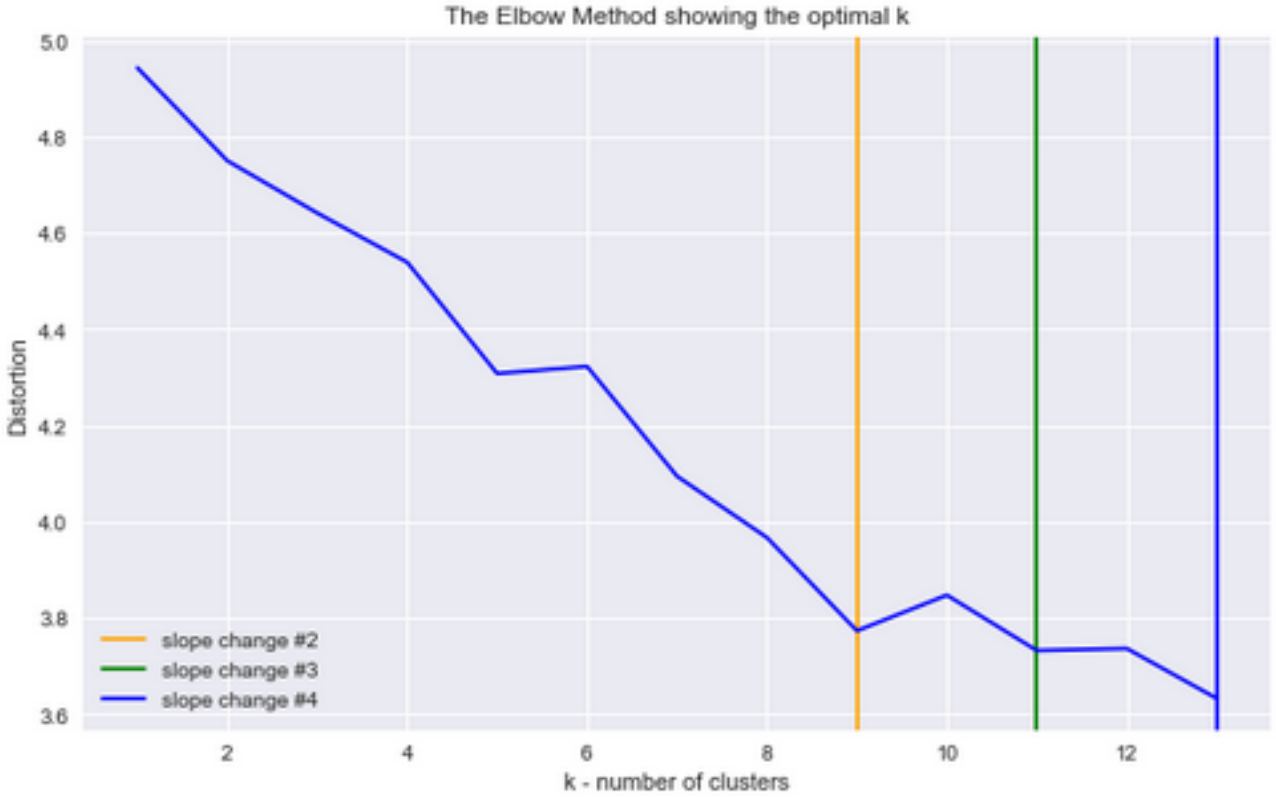
-

- Чуствителен к начальному приближению. Могут быть ситуации, когда сойдемся к локальному экстремуму.
- Число компонент К является гиперпараметром

Метод кластеризации: K-Means

Метод k-Means –алгоритм обучения без учителя, который группирует элементы в k кластеров так, чтобы расстояние до центра кластера было минимальным.

1. Назначение числа кластеров (k), на которое должны быть разбиты данные
2. Случайная инициализация центров кластеров
3. Присвоение номера группы каждому элементу по самому близкому центроиду
4. Пересчет координат центров кластеров
5. Повторение 3 и 4 шагов до тех пор, пока назначения кластеров не прекращают изменяться или не достигнуто заданное число итераций



Расстояние до центра кластера рассчитывается по формуле Евклидова расстояния:

$$W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \overline{x}_k||_2^2$$

Идея метода заключается в минимизации расстояния от каждой точки кластера до его центра:

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} W(C_k)$$

\+

Простота в использовании
Скорость

\-

Неустойчивость кластерной структуры итеративным построениям

Методы и индексы качества кластеризации

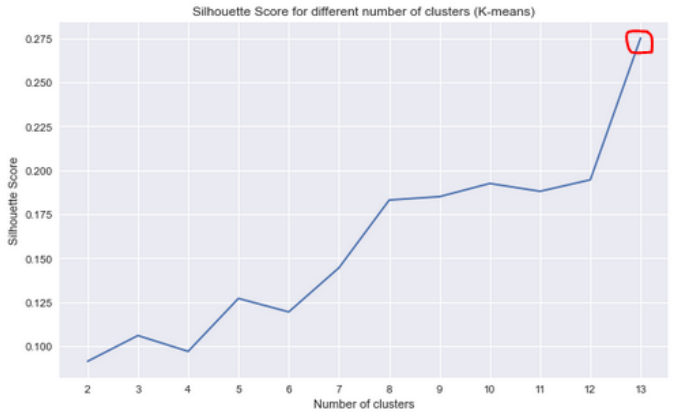
Задачи:
- Определение оптимального кол-ва кластеров для каждого метода
- Определение наилучшего метода кластеризации для набора данных

Силуэт

$$Sil(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k), b(x_i, c_k)\}},$$

Показывает, насколько объект похож на свой кластер по сравнению с другими кластерами.

Чем ближе значение Силуэта к 1, тем лучше кластеризация.

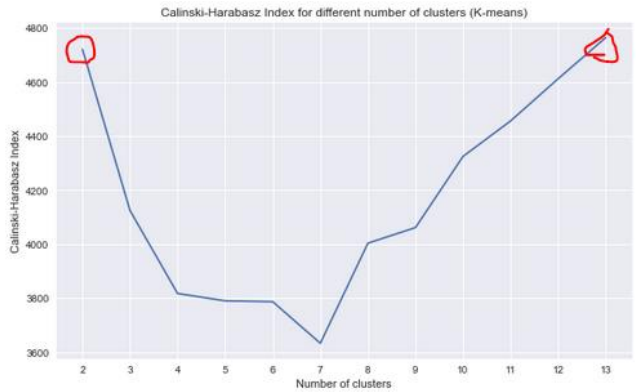


Индекс Калинского-Харабаза

$$CH(C) = \frac{N - K}{K - 1} \cdot \frac{\sum_{c_k \in C} |c_k| \cdot \|\overline{c_k} - \overline{X}\|}{\sum_{c_k \in C} \sum_{x_i \in c_k} \|x_i - \overline{c_k}\|}$$

Компактность основана на расстоянии от точек кластера до их центроидов, а разделимость - на расстоянии от центроид кластеров до глобального центроида.

Чем больше значение индекса, тем лучше кластеризация.

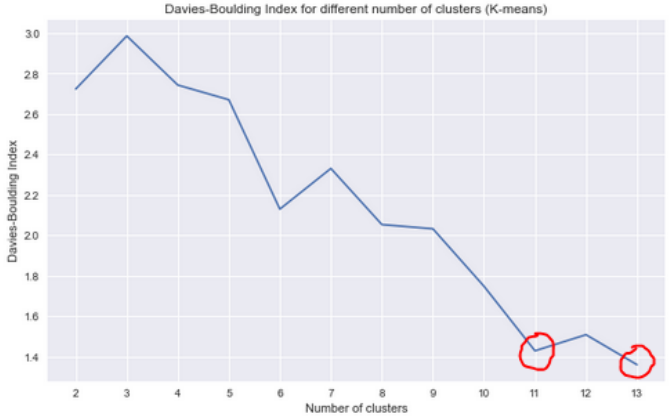


Индекс Дэвиса-Болдуина

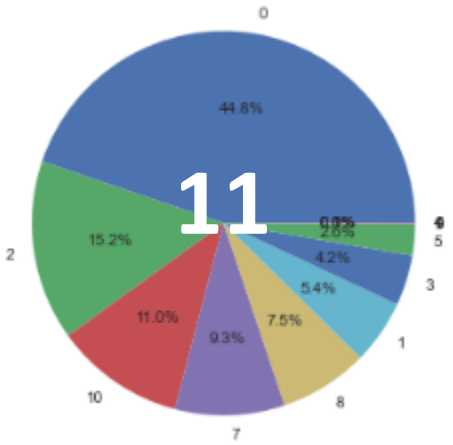
$$DB(C) = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{S(c_k) + S(c_l)}{\|\overline{c_k} - \overline{c_l}\|} \right\},$$

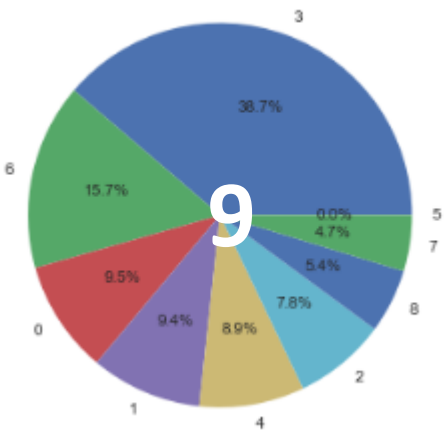
Вычисляет компактность как расстояние от объектов кластера до их центроидов, а отделимость - как расстояние между центроидами.

Чем меньше значение индекса, тем лучше кластеризация.

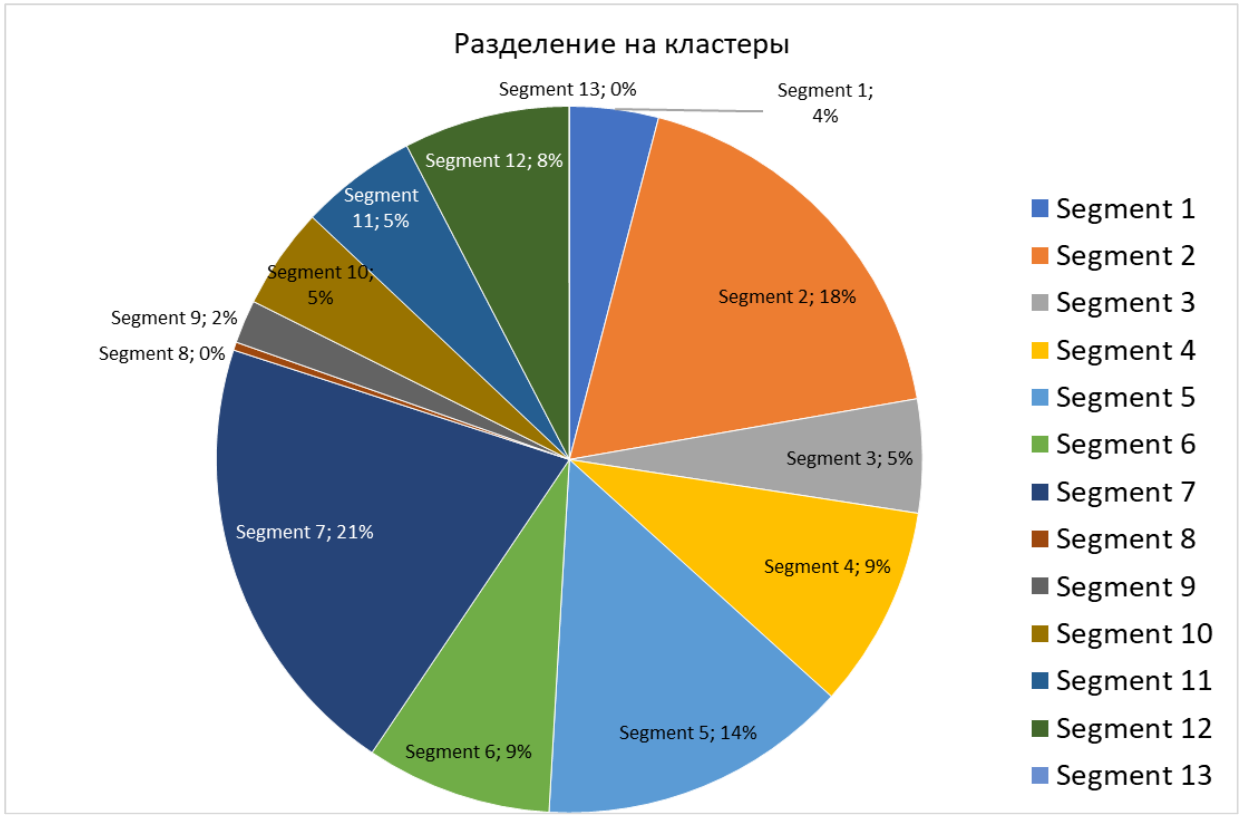


Разделение на кластеры






```
Segment 1:   3480
Segment 2:   15547
Segment 3:   4436
Segment 4:   7919
Segment 5:   12164
Segment 6:   7288
Segment 7:   17594
Segment 8:   322
Segment 9:   1704
Segment 10:   3976
Segment 11:   4580
Segment 12:   6509
Segment 13:   3
df:   85522
```



13 Кластеров

3 480 клиентов (4%)
Ср. чек: 2 758 ₽
Ср. кол-во тов. чеке: 1,12
Осн. кат.: Товары для дома
Выручка: 11 077 884 ₽

15 547 клиентов (18%)
Ср. чек: 7 661 ₽
Ср. кол-во тов. чеке: 1,02
Осн. кат.: Аксессуары, цифровая техника, автотовары, услуги, хобби
Выручка: 125 169 160 ₽

4 436 клиентов (5%)
Ср. чек: 8 948 ₽
Ср. кол-во тов. чеке: 1
Осн. кат.: Автотовары, услуги
Выручка: 43 215 747 ₽

7 919 клиентов (9%)
Ср. чек: 6 400 ₽
Ср. кол-во тов. чеке: 1,02
Осн. кат.: Сделай сам
Выручка: 53 455 755 ₽

12 164 клиентов (14%)
Ср. чек: 18 377 ₽
Ср. кол-во тов. чеке: 1
Осн. кат.: Крупная бытовая техника
Выручка: 237 197 910 ₽

7 288 клиентов (9%)
Ср. чек: 6 912 ₽
Ср. кол-во тов. чеке: 1,07
Осн. кат.: Спорт и активный отдых
Выручка: 52 818 775 ₽

17 594 клиентов (21%)
Ср. чек: 5 019 ₽
Ср. кол-во тов. чеке: 1,01
Осн. кат.: Мелкая бытовая техника
Выручка: 89 843 042 ₽

322 клиентов (0,004%)
Ср. чек: 3 621 ₽
Ср. кол-во тов. чеке: 1,01
Осн. кат.: Игры, софт и развлечения
Выручка: 1 306 970 ₽

1 704 клиентов (2%)
Ср. чек: 7 481 ₽
Ср. кол-во тов. чеке: 1
Осн. кат.: Климат
Выручка: 12 960 295 ₽

3 976 клиентов (5%)
Ср. чек: 4 369 ₽
Ср. кол-во тов. чеке: 1,01
Осн. кат.: Мебель
Выручка: 17 607 185 ₽

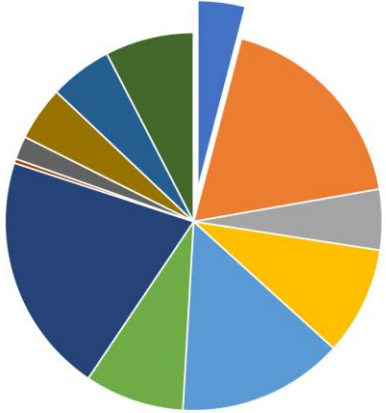
4 580 клиентов (5%)
Ср. чек: 15 431 ₽
Ср. кол-во тов. чеке: 1,01
Осн. кат.: ТВ-аудио
Выручка: 73 925 690 ₽

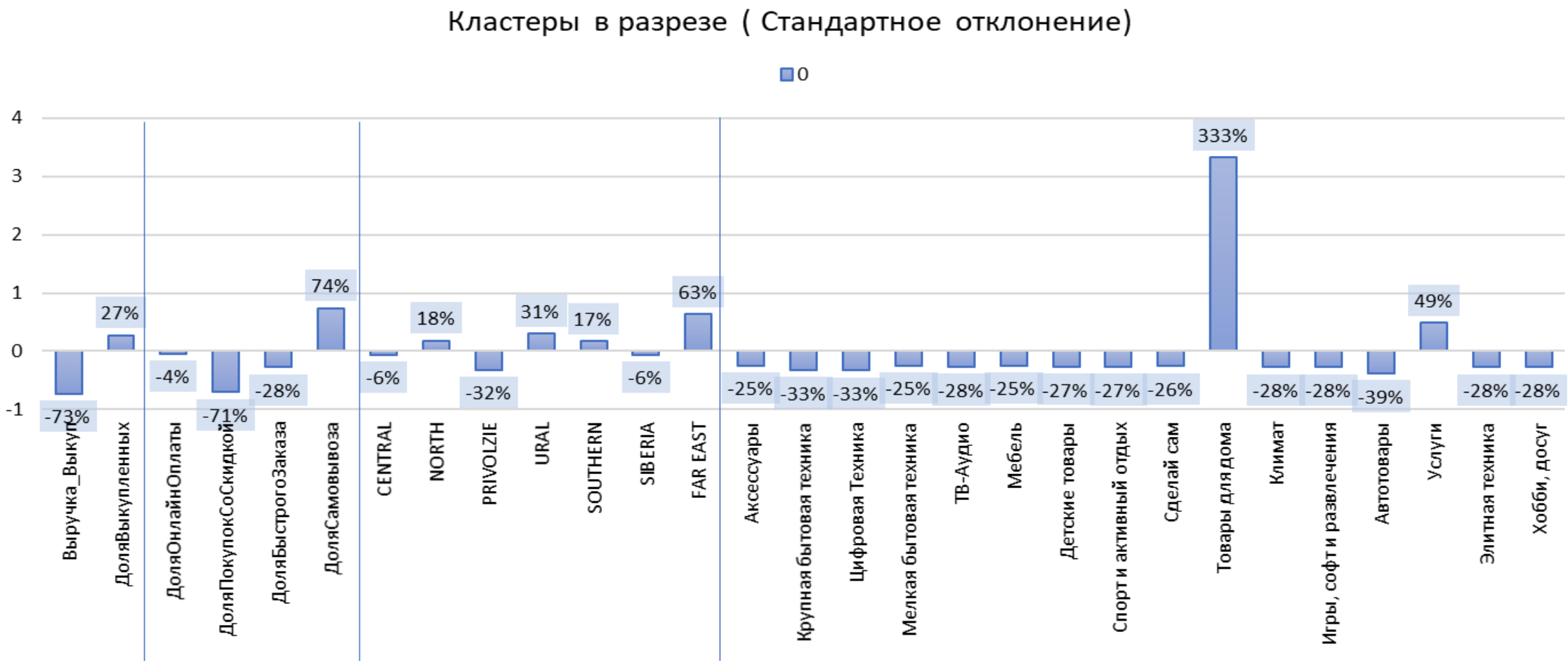
6 509 клиентов (8%)
Ср. чек: 6 084 ₽
Ср. кол-во тов. чеке: 1,01
Осн. кат.: Детские товары
Выручка: 41 281 931 ₽

3 клиентов (0%)
Ср. чек: 46 626 ₽
Ср. кол-во тов. чеке: 1
Осн. кат.: Элитная техника
Выручка: 139 878 ₽

1 Кластер

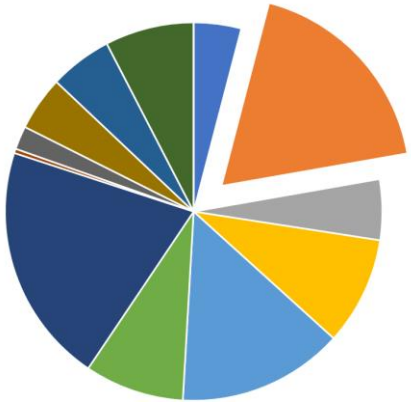
3 480 клиентов (4%)

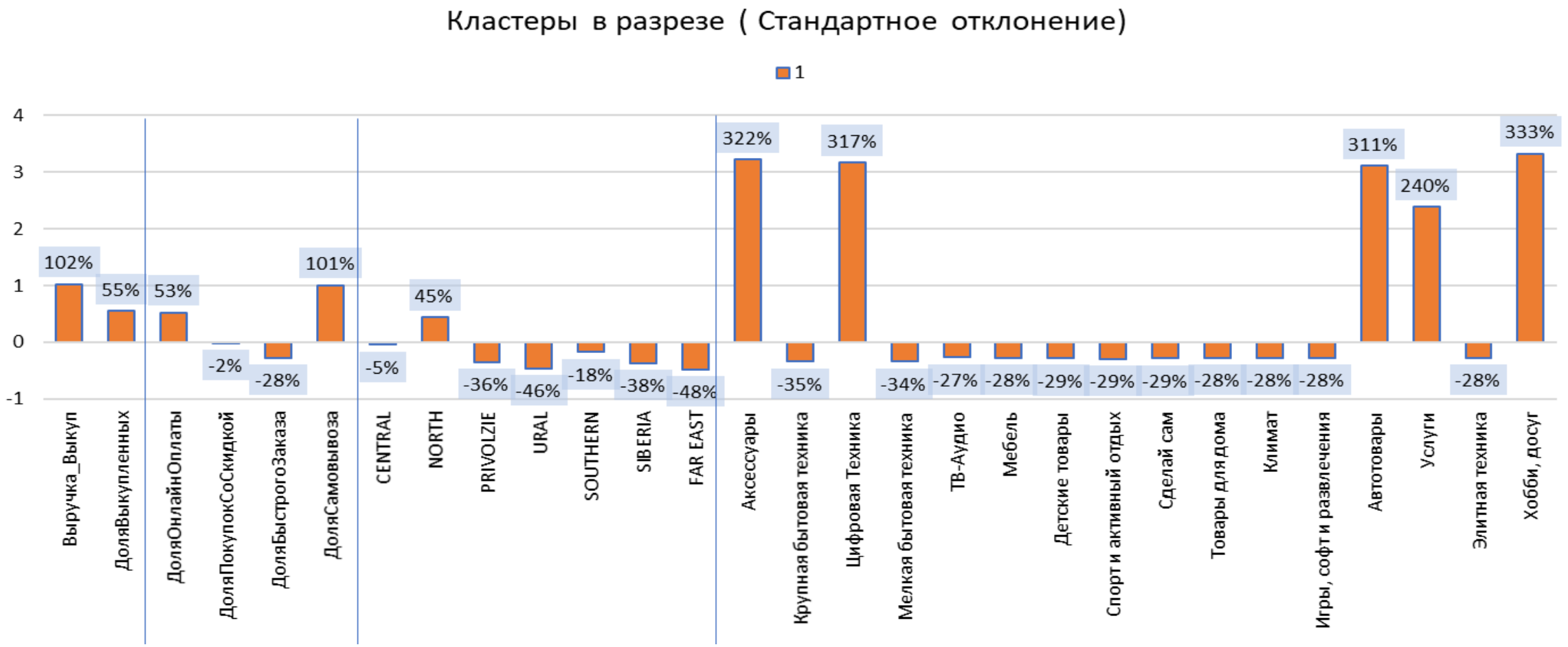




2 Кластер

15 547 клиентов (18%)





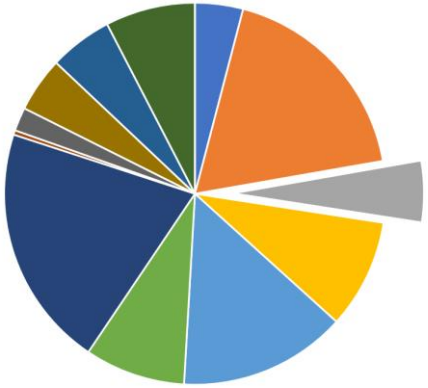
Благодарим за внимание!

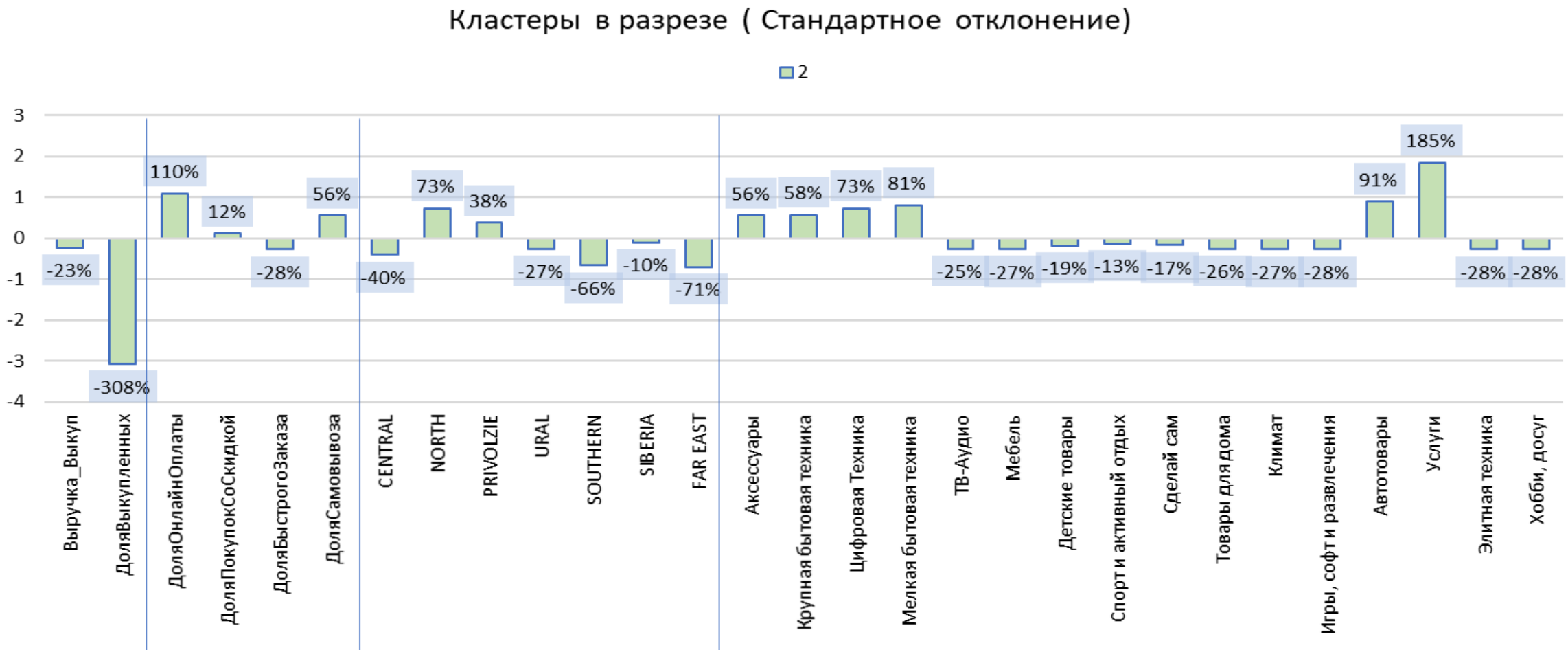
Список источников

1. http://statistica.ru/
2. https://algowiki-project.org/
3. https://ru.wikipedia.org/
4. https://ranalytics.github.io/
5. https://learn.innopolis.university/
6. https://datstat.ru/

3 Кластер

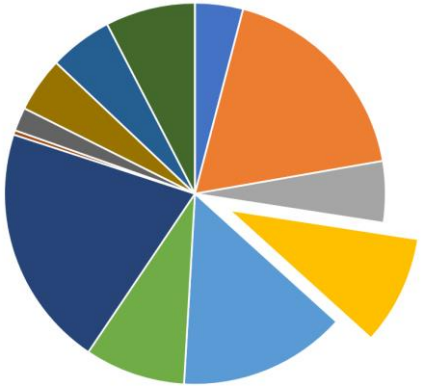
4 436 клиентов (5%)

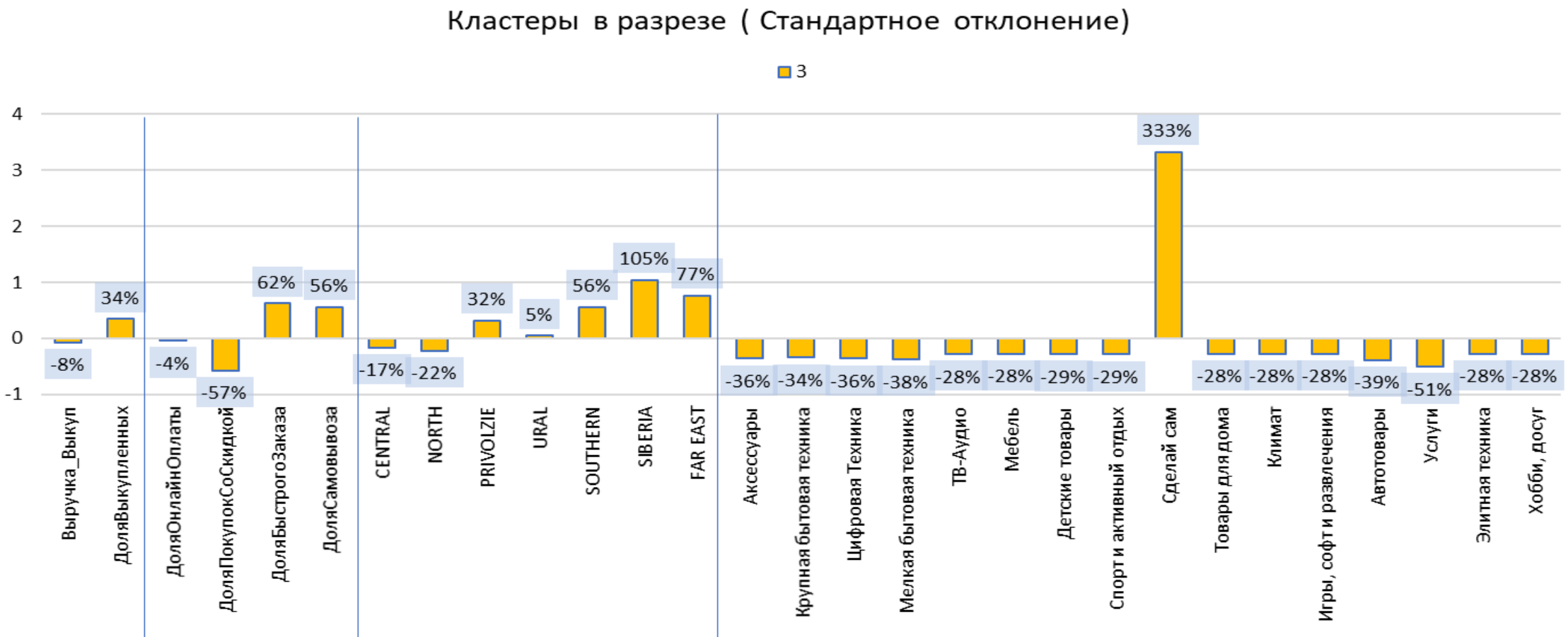




4 Кластер

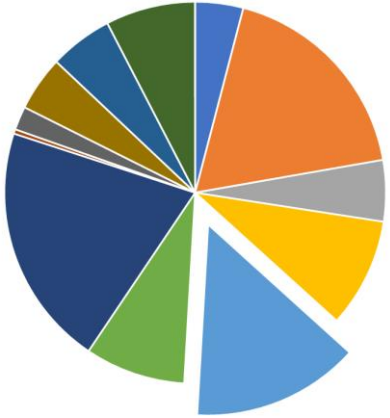
7 919 клиентов (9%)

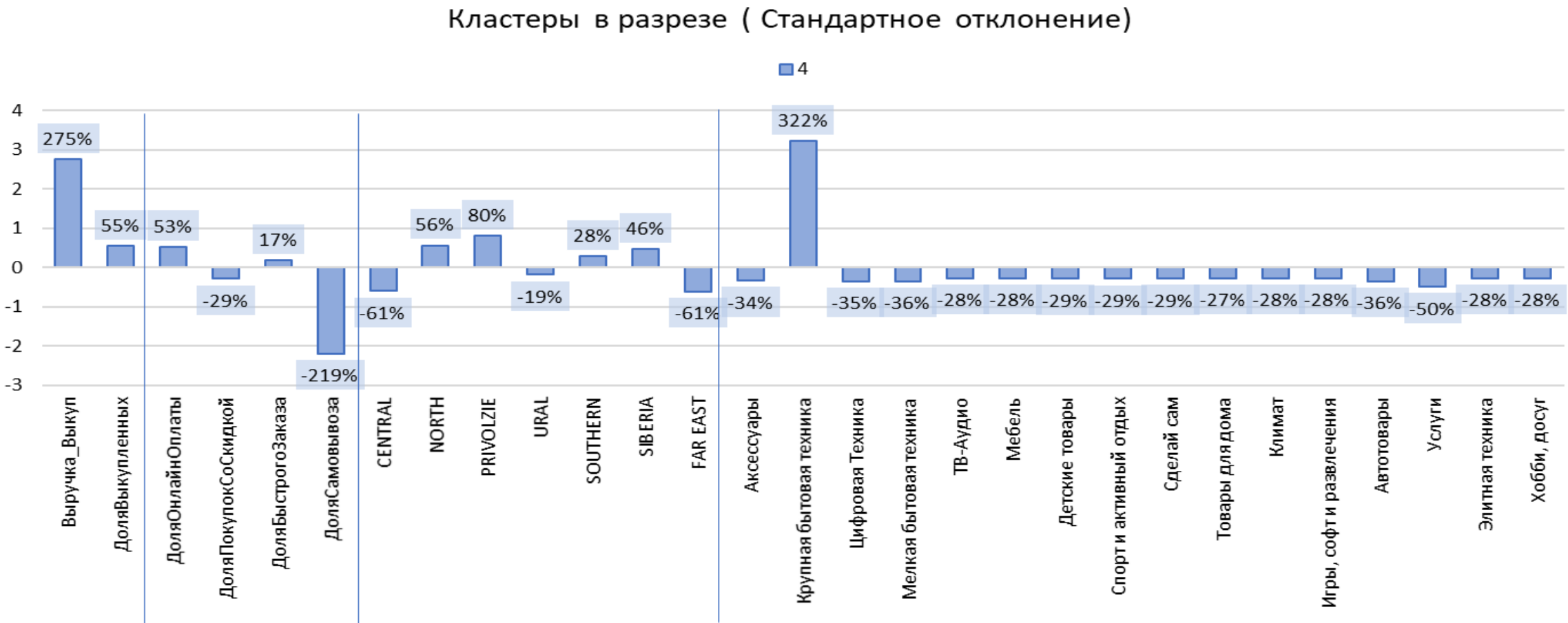




5 Кластер

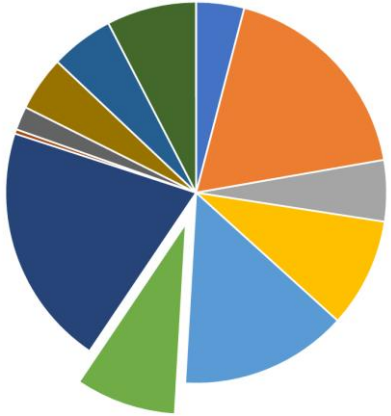
12 164 клиентов (14%)

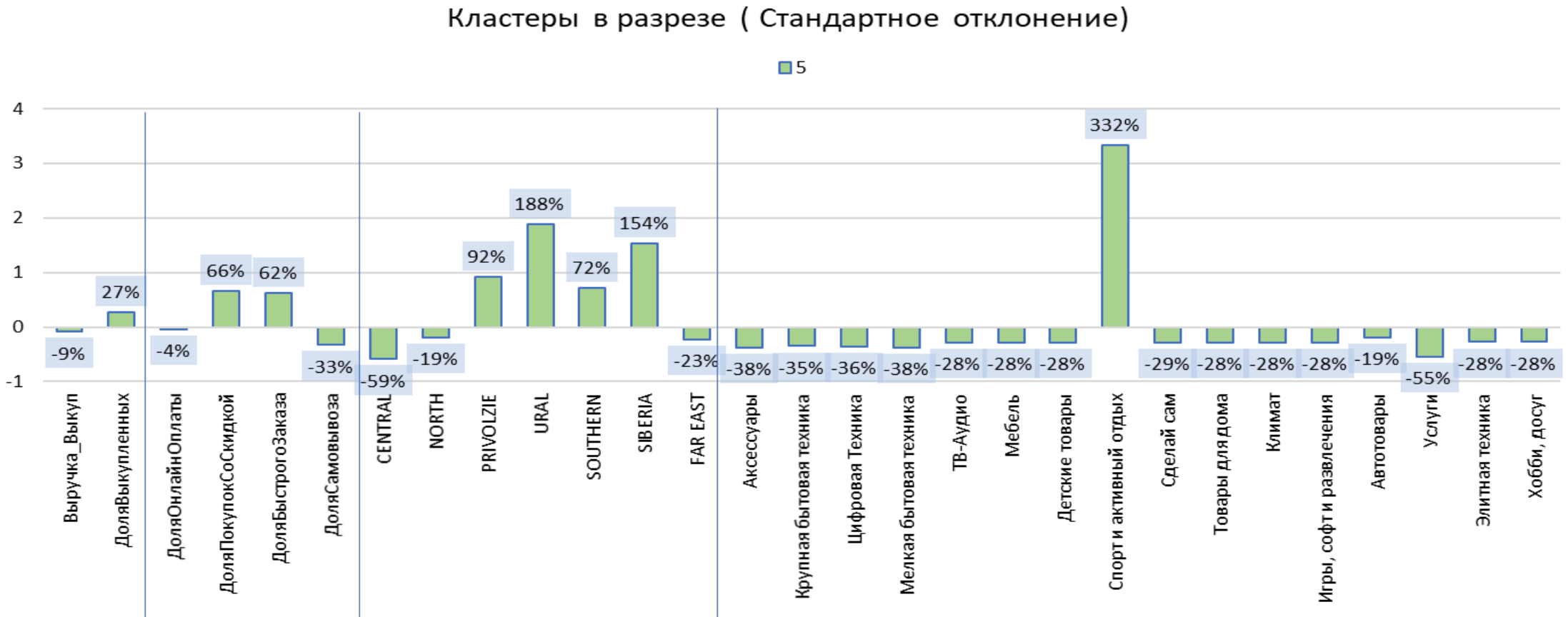




6 Кластер

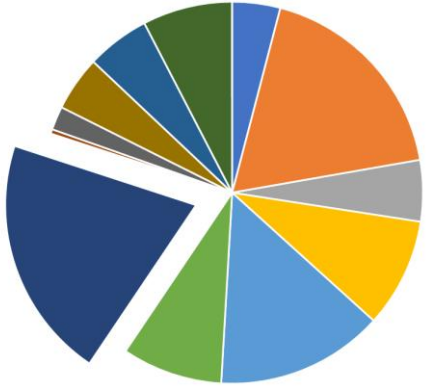
7 288 клиентов (9%)

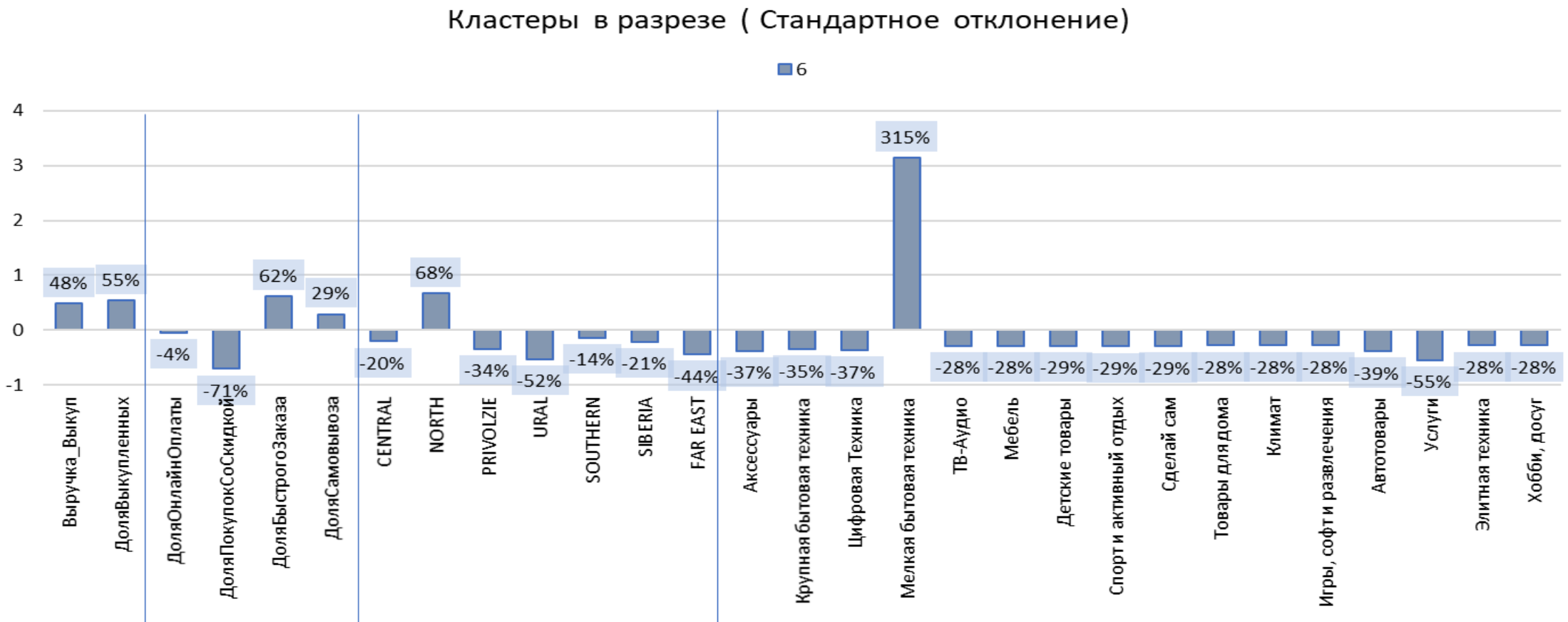




7 Кластер

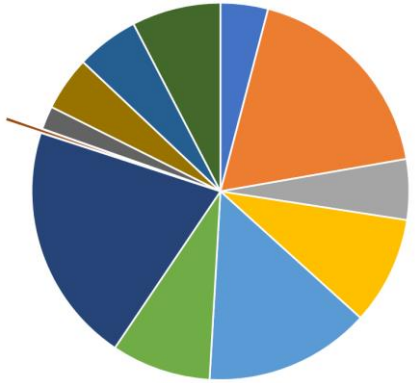
17 594 клиентов (21%)

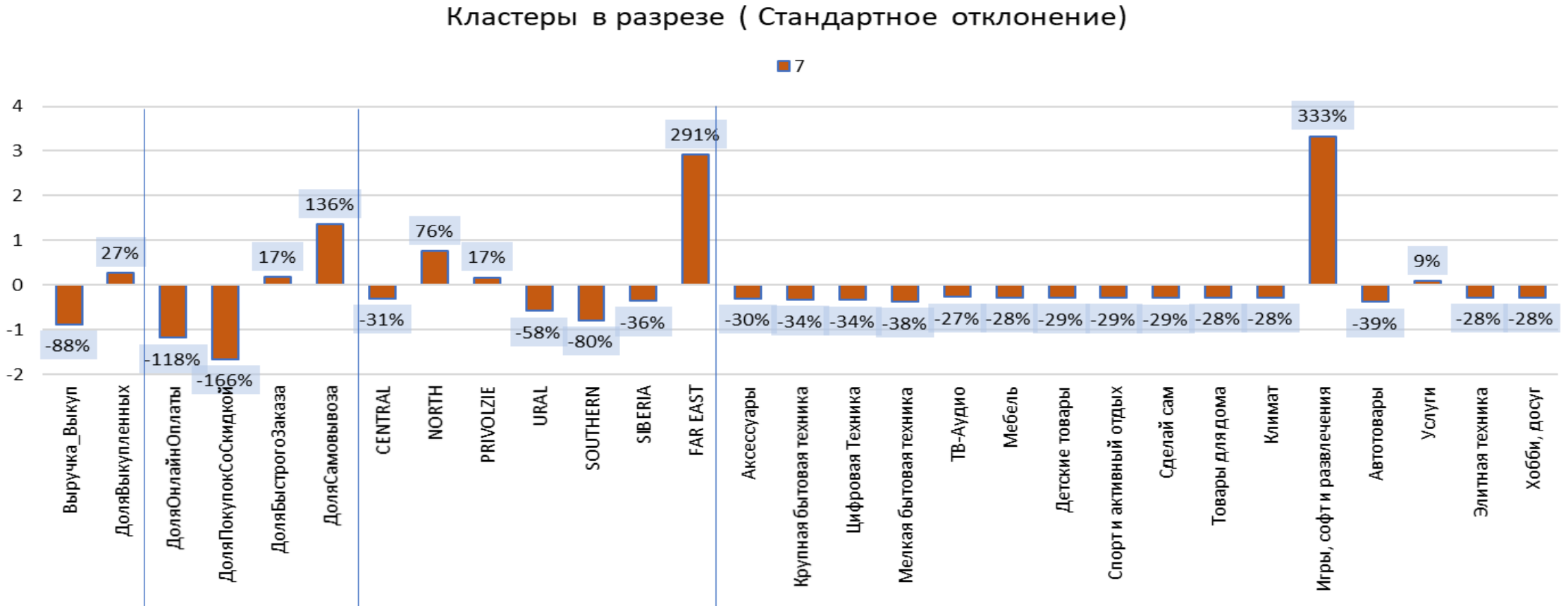




8 Кластер

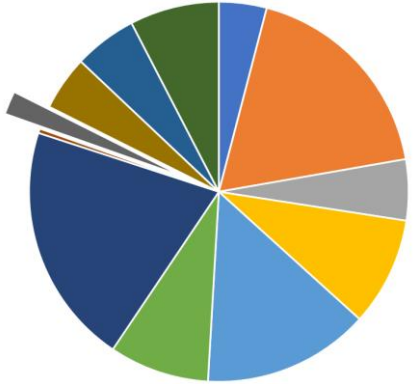
322 клиентов (0,004%)

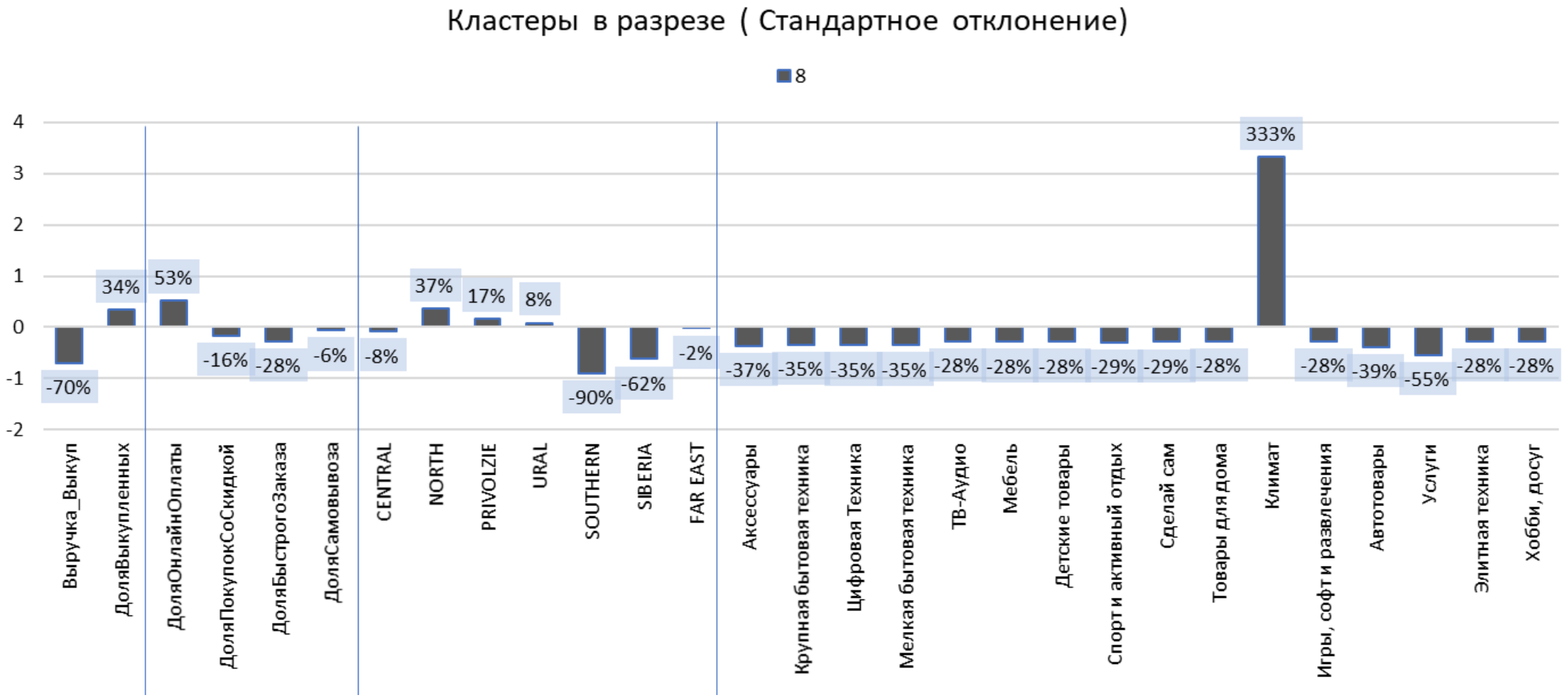




9 Кластер

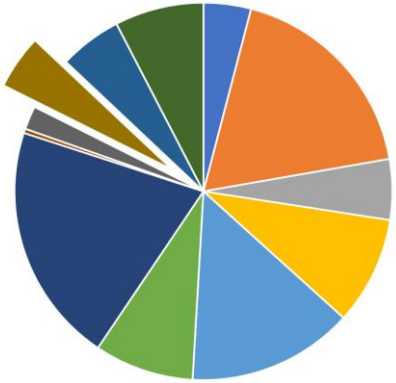
1 704 клиентов (2%)

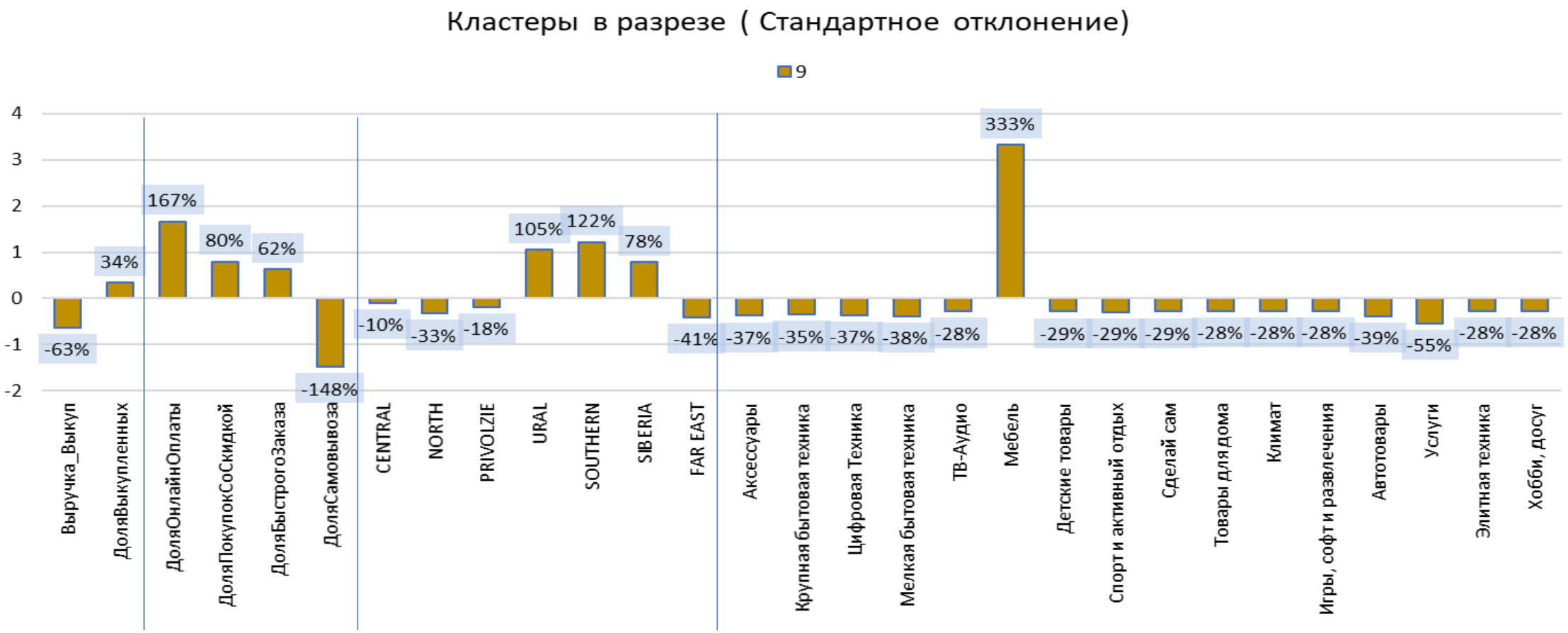




10 Кластер

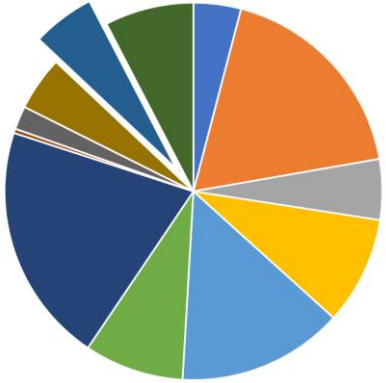
3 976 клиентов (5%)

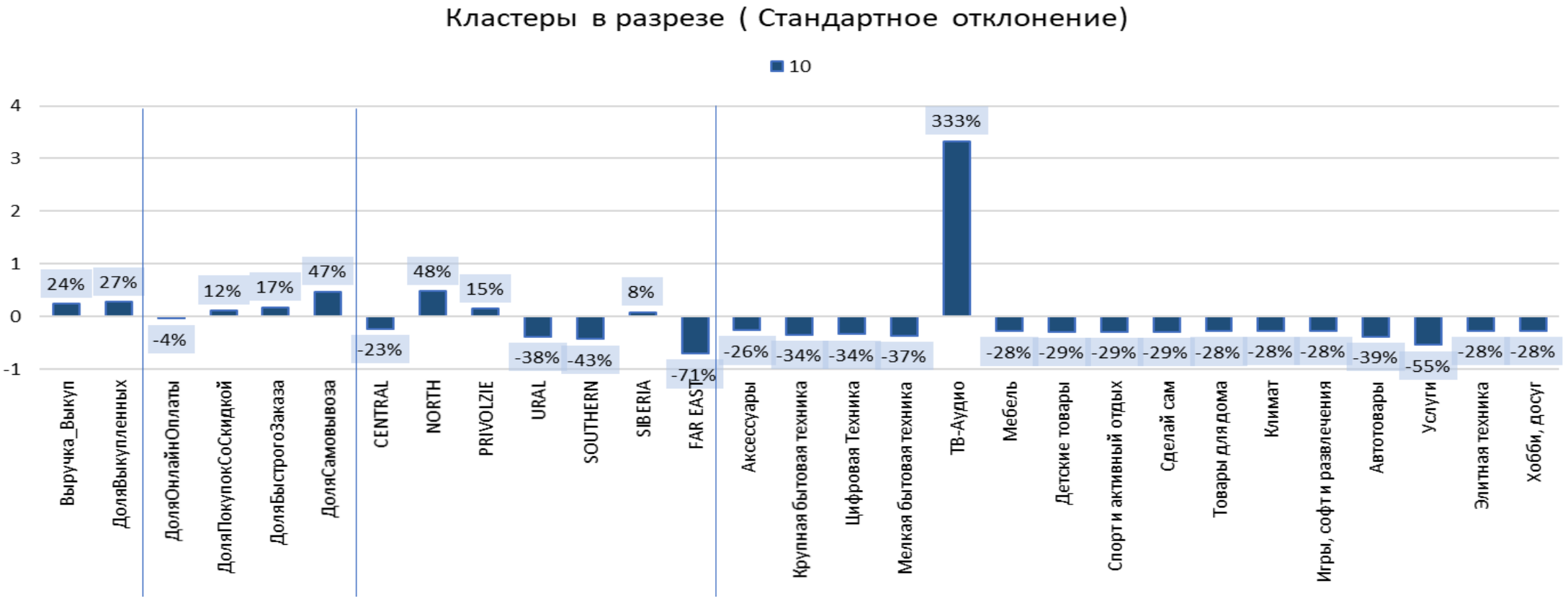




11 Кластер

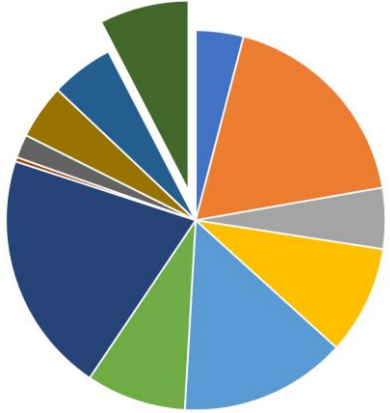
4 580 клиентов (5%)

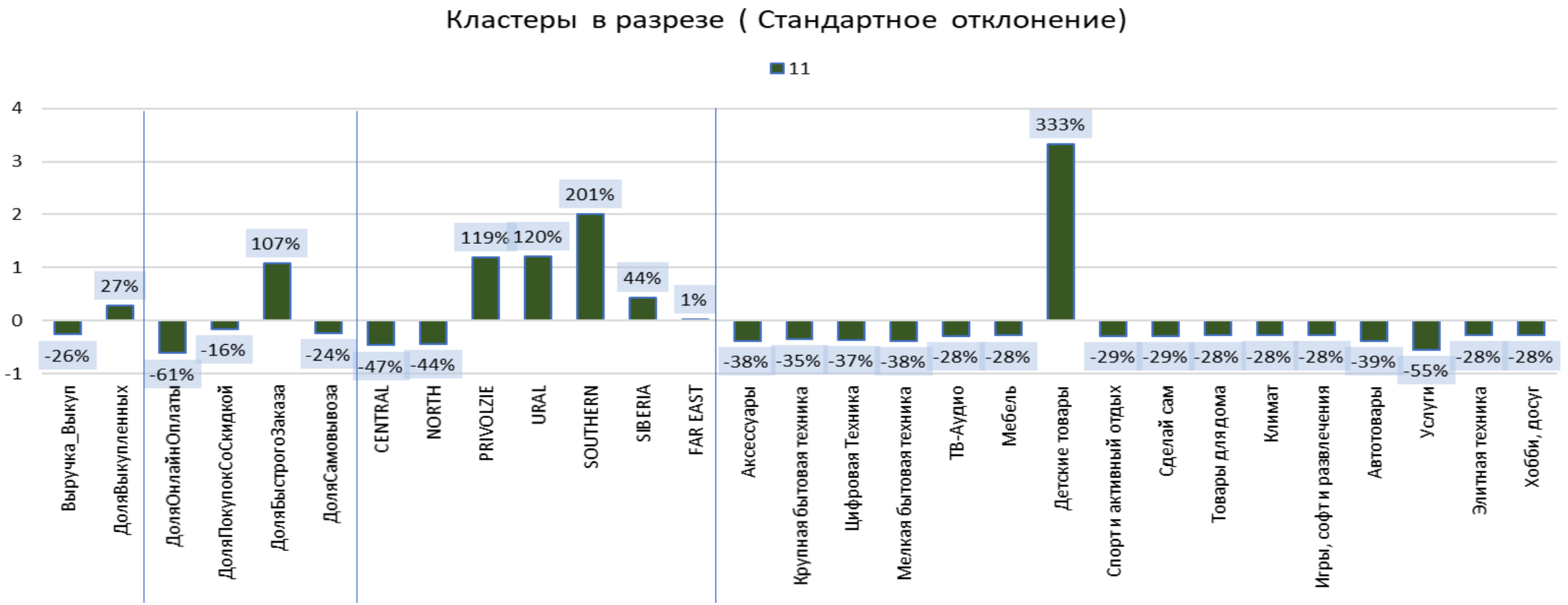




12 Кластер

6 509 клиентов (8%)





13 Кластер

3 клиентов (0%)