Кластеризация пользователей

Подготовил Дитятев Иван

Intro

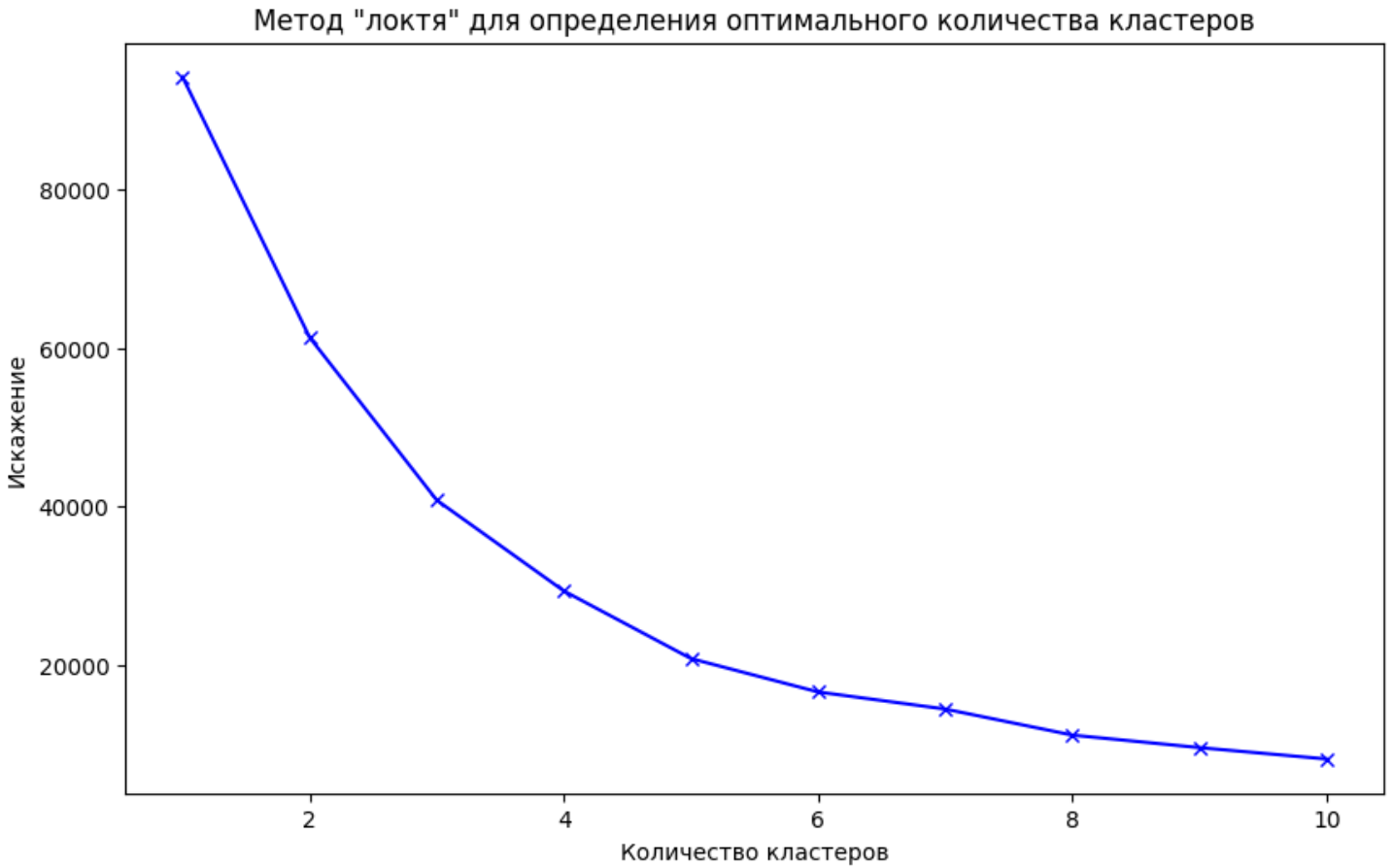
Цель:

- Разделить пользователей на кластеры.
- Предлагаемые кластеры:
- - активный донор,
- - спящий донор,
- - реактивированный донор,
- - потерянный.

Задачи:

- - Предобработка данных.
- - Определение оптимального числа кластеров.
- - Разделение пользователей на кластеры.
- - Определение четких правил, как отнести пользователя к тому или иному кластеру.
- - Исследовательский анализ данных.
- - Создание презентации.

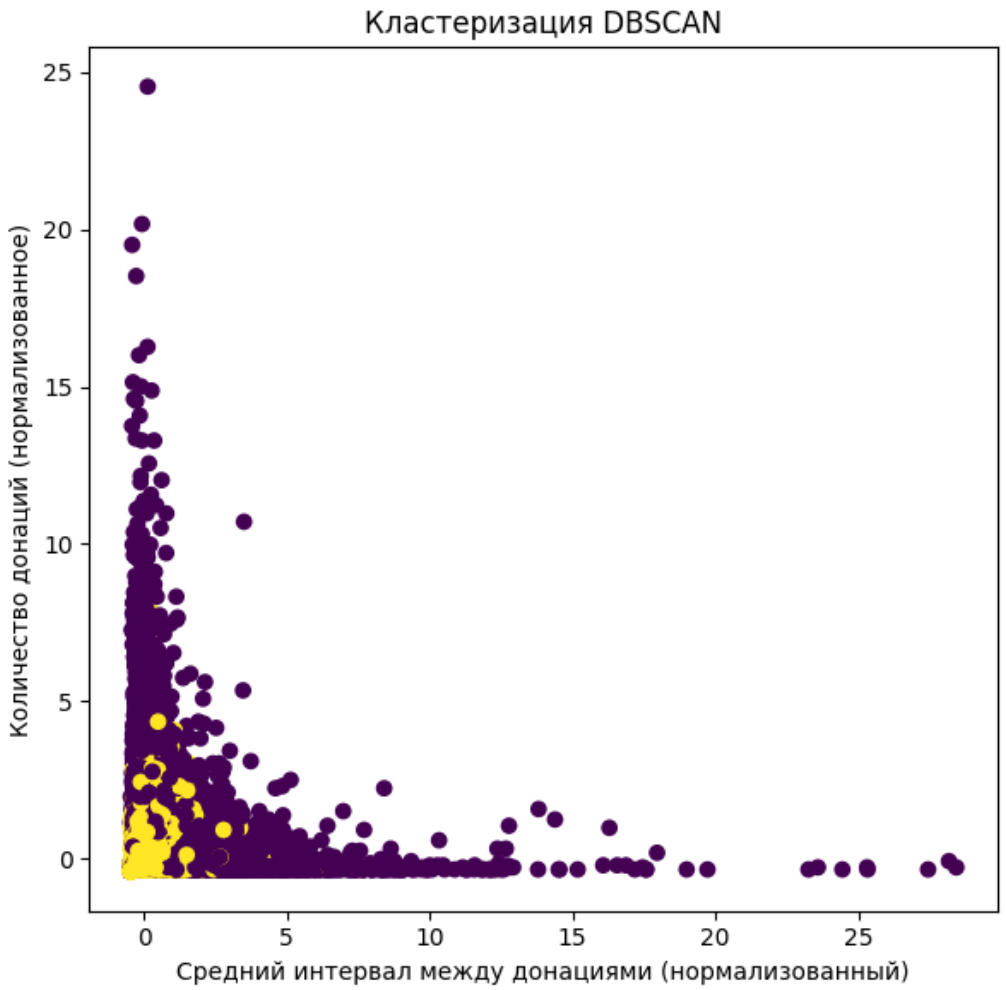
Оптимальное кол-во кластеров



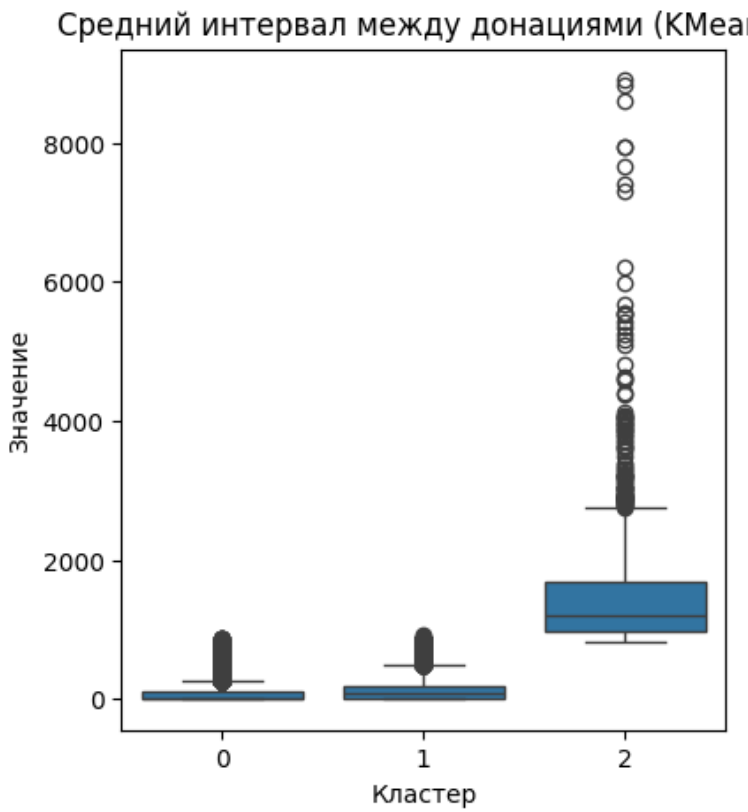
Оптимальное кол-во = 3 кластера

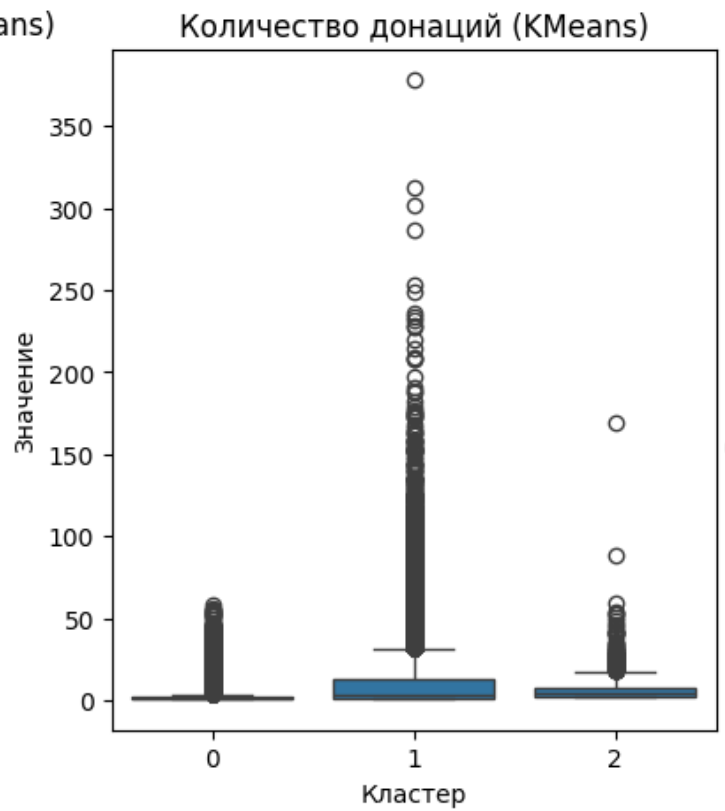
Кластеризация

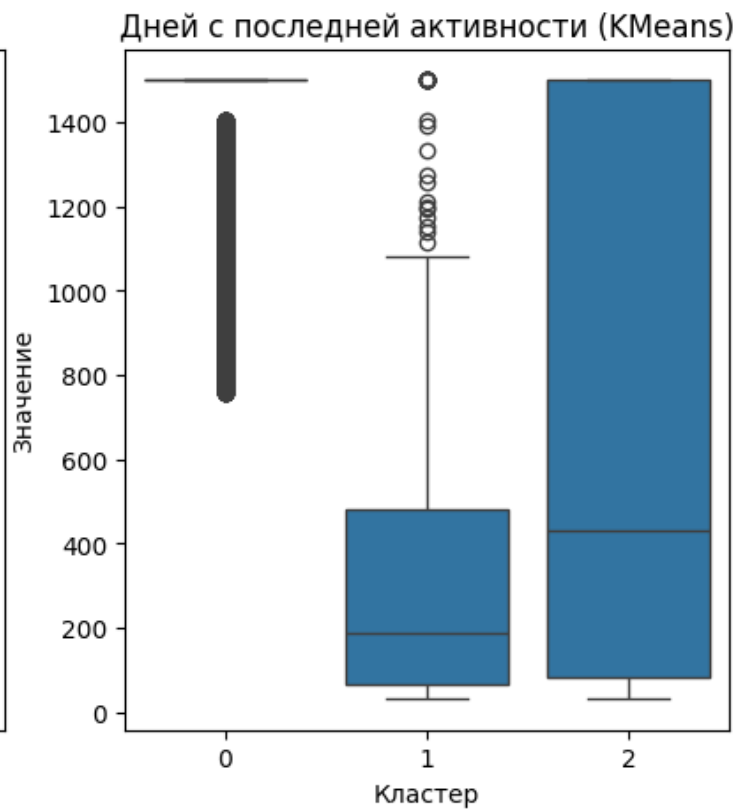
Эта модель выглядит лучше



Распределение пользователей по кластерам







Распределение пользователей по кластерам

Активный донор (0):
- Количество донаций: <= 3.5
- Средний интервал между донациями: <= 252.5
- Дни с последней активности: <= 1500
- Причина: Эти доноры регулярно сдают кровь, имеют небольшие интервалы между донациями и относительно недавно были активны.

Спящий донор (1):
- Количество донаций: <= 31
- Средний интервал между донациями: <= 490
- Дни с последней активности: <= 1103.5
- Причина: Эти доноры не сдавали кровь в течение длительного времени, но могут вернуться к донорству. У них большее количество донаций и средние промежутки между ними.

Реактивированный донор (2):
- Количество донаций: <= 17
- Средний интервал между донациями: <= 2752.625
- Дни с последней активности: <= 3622.5
- Причина: Эти доноры снова начали сдавать кровь после длительного перерыва. У них умеренное количество донаций и длинные промежутки между ними.

Потерянный донор (3):
- Количество донаций: > 31 или Средний интервал между донациями > 2752.625 или Дни с последней активности > 3622.5
- Причина: Эти доноры не сдавали кровь в течение очень длительного времени и, вероятно, не вернутся к донорству. У них очень длинные промежутки между донациями и небольшое количество донаций.

Определение правил распределения по кластерам – ВЫВОД

		Значение									
		count	mean	std	min	25%	50%	75%	max	lb	ub
Кластер_kmeans	Показатель										
0	Дни с последней активности	22531.0	1427.619546	190.176261	759.00	1500.000000	1500.000000	1500.0	1500.000000	1500.0	1500.000
	Количество донаций	22531.0	2.627313	4.171850	1.00	1.000000	1.000000	2.0	58.000000	0.0	3.500
	Средний интервал между донациями	22531.0	76.276041	142.640807	0.00	0.000000	0.000000	101.0	867.000000	0.0	252.500
1	Дни с последней активности	23483.0	283.141549	242.352454	32.00	66.000000	190.000000	481.0	1500.000000	0.0	1103.500
	Количество донаций	23483.0	11.418643	19.990569	1.00	1.000000	3.000000	13.0	378.000000	0.0	31.000
	Средний интервал между донациями	23483.0	134.532684	167.199113	0.00	0.000000	84.000000	196.0	921.101449	0.0	490.000
2	Дни с последней активности	1120.0	664.721429	603.266966	32.00	85.000000	431.500000	1500.0	1500.000000	0.0	3622.500
	Количество донаций	1120.0	7.427679	9.815280	2.00	2.000000	4.000000	8.0	169.000000	0.0	17.000
	Средний интервал между донациями	1120.0	1544.684856	971.490374	811.25	982.416667	1209.666667	1690.5	8898.000000	0.0	2752.625

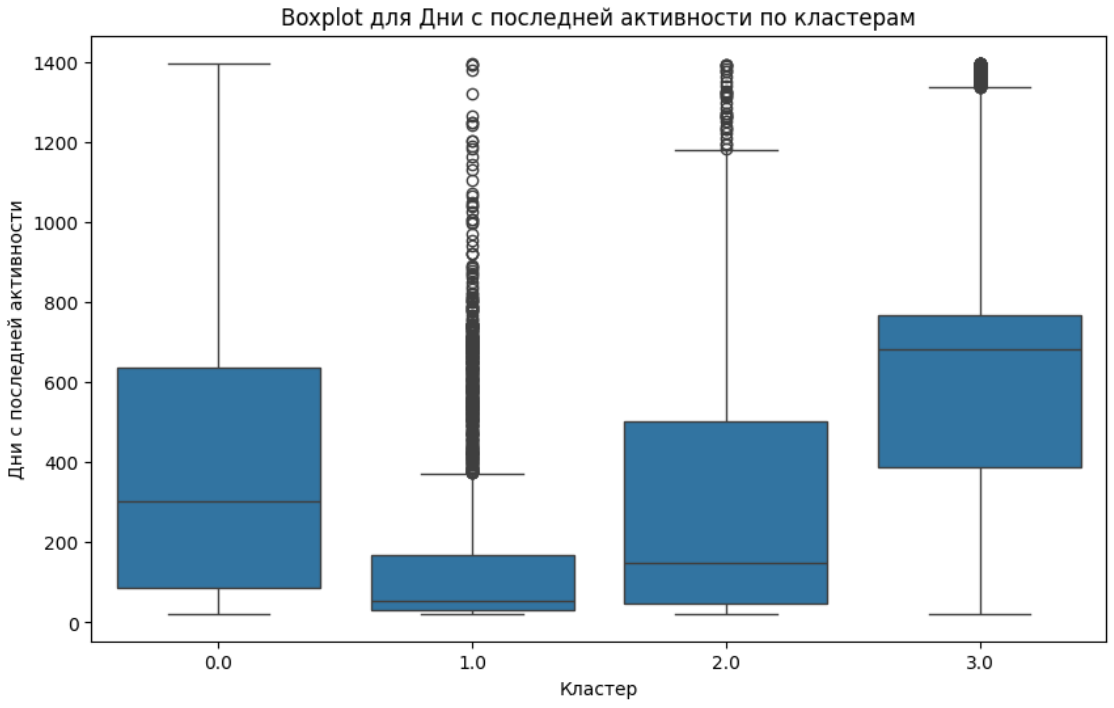
Пример запроса:

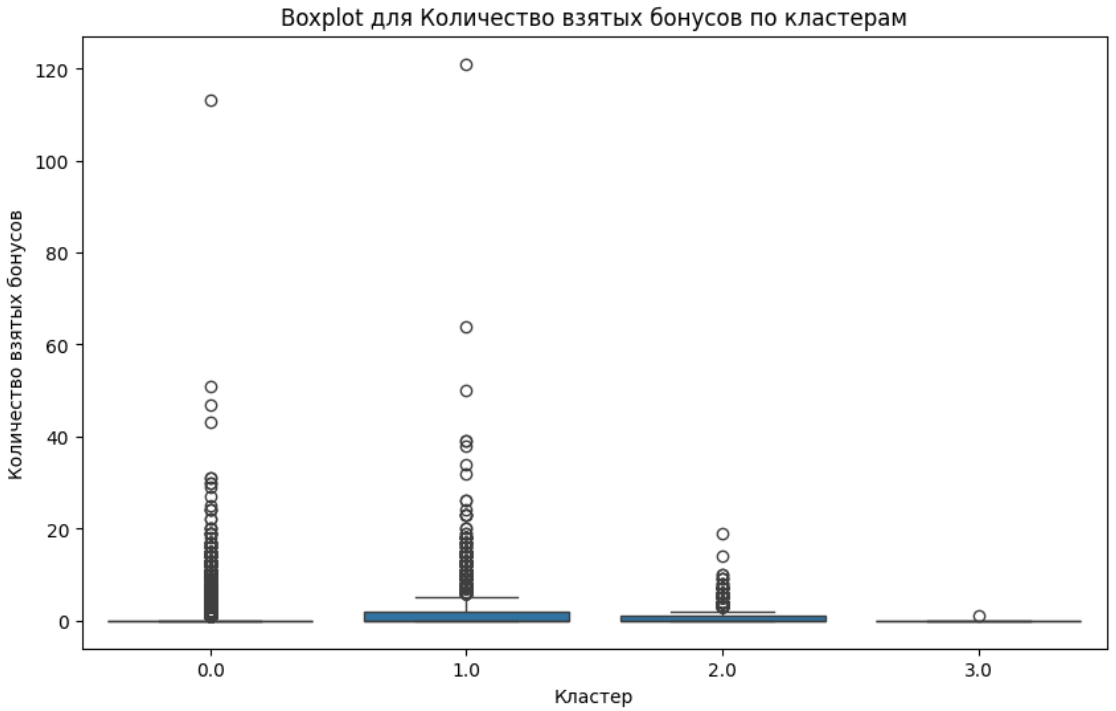
```sql
SELECT *,
    CASE
        WHEN Количество_донаций <= 3.5 AND Средний_интервал_между_донациями <= 252.5 AND Дни_с_последней_активности <= 1500 THEN
'Активный донор'
        WHEN Количество_донаций <= 31 AND Средний_интервал_между_донациями <= 490 AND Дни_с_последней_активности <= 1103.5 THEN 'Спящий
донор'
        WHEN Количество_донаций <= 17 AND Средний_интервал_между_донациями <= 2752.625 AND Дни_с_последней_активности <= 3622.5 THEN
'Реактивированный донор'
        ELSE 'Потерянный донор'
    END AS Кластер
FROM some_table;
```

Проверка наличия статически значимых различий между кластерами t-test/ Mann-Whitney U Test (для числовых колонок), chi^2 (для категориальных).

- Проверка показала наличие статистически значимых различий между кластерами.

Некоторые прикольные инсайты с исследовательского анализа





Некоторые прикольные инсайты с исследовательского анализа