Структура клиентской базы и влияние на первые месяцы покупок

Данное исследование сделано на основании датасета:

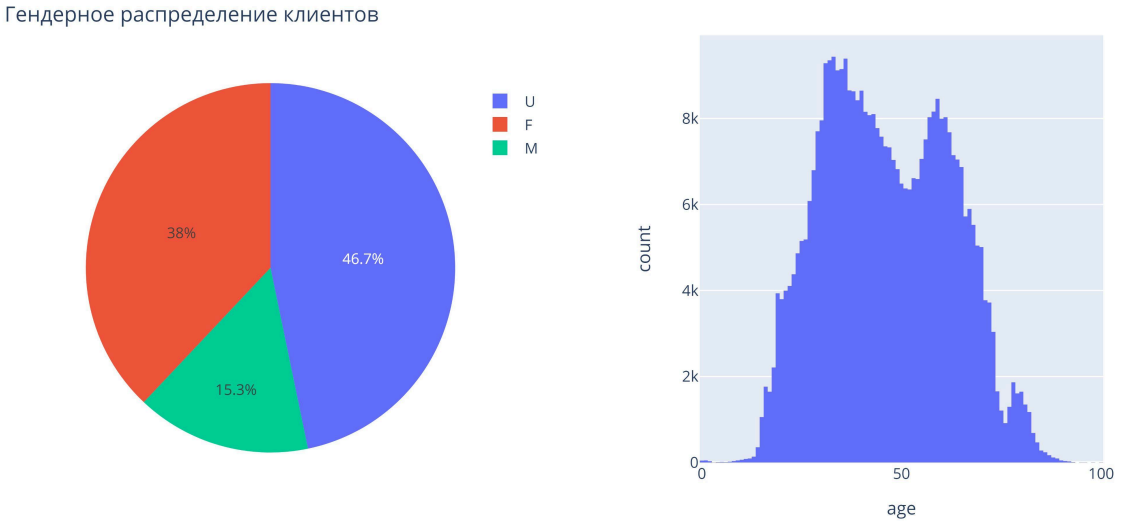
https://www.kaggle.com/datasets/mvyurchenko/x5-retail-hero/

Он был использован в соревновании, посвященном созданию рекомендательных систем, однако попробуем посмотреть на него с точки зрения аналитики данных, для этого свяжем данные о клиентах, их гендерной принадлежности, возрасте и дате приобретения карты с данными о покупках.

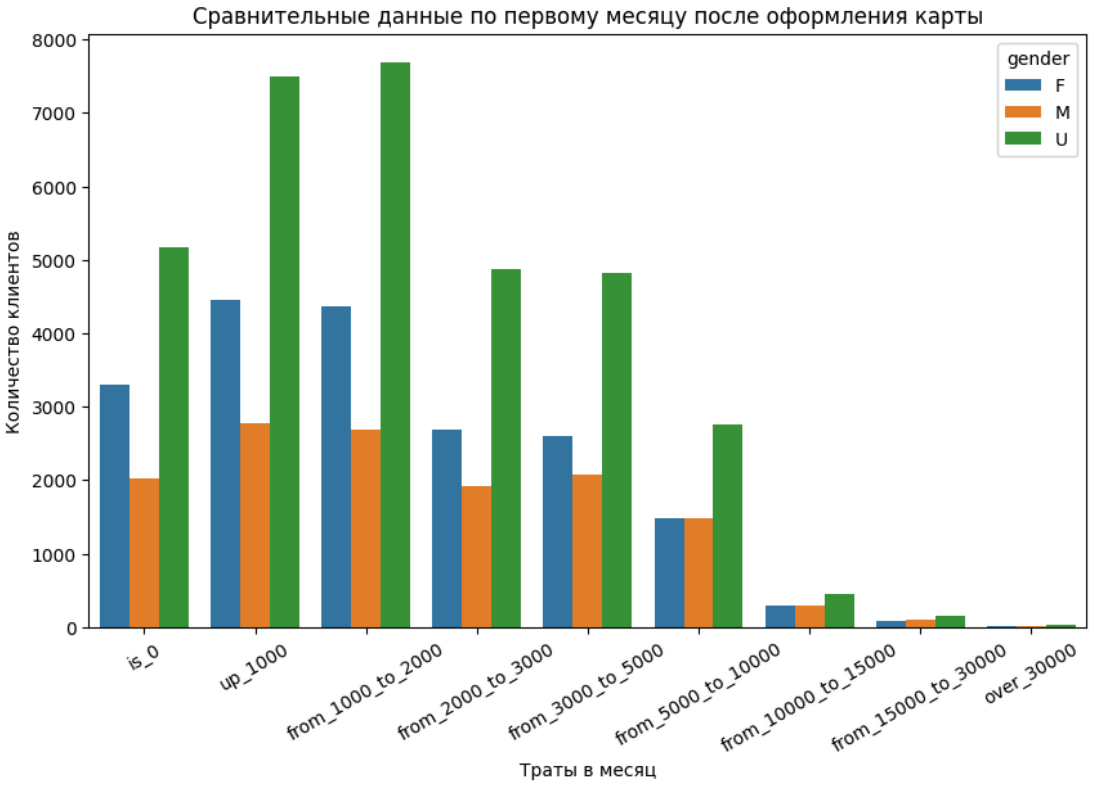
Во-первых, рассмотрим данные о приобретении карт лояльности.

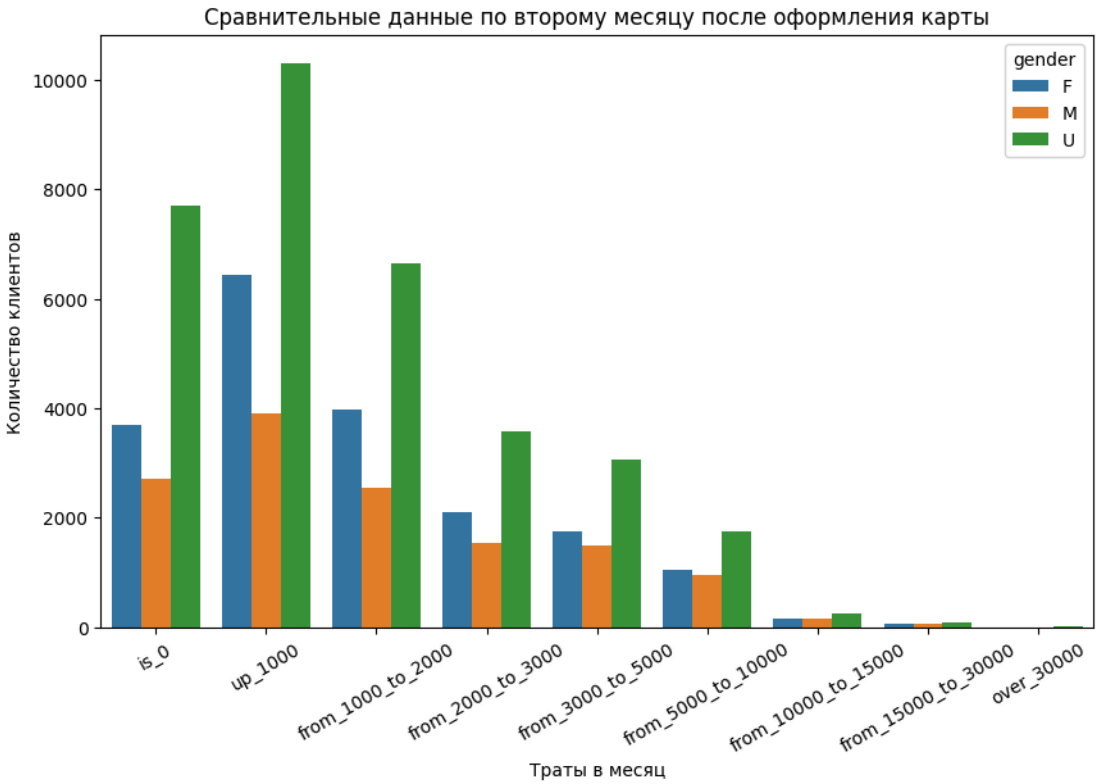
Нетрудно заметить, что начало нового года является точкой структурного изменения. Более того, в рамках 2017 года можем найти другие точки структурных изменений.

Во-первых, рассмотрим половозрастной состав клиентов. Заметим, что клиентам разрешено не указывать пол, поэтому большая доля клиентов имеет неидентифицированный пол.

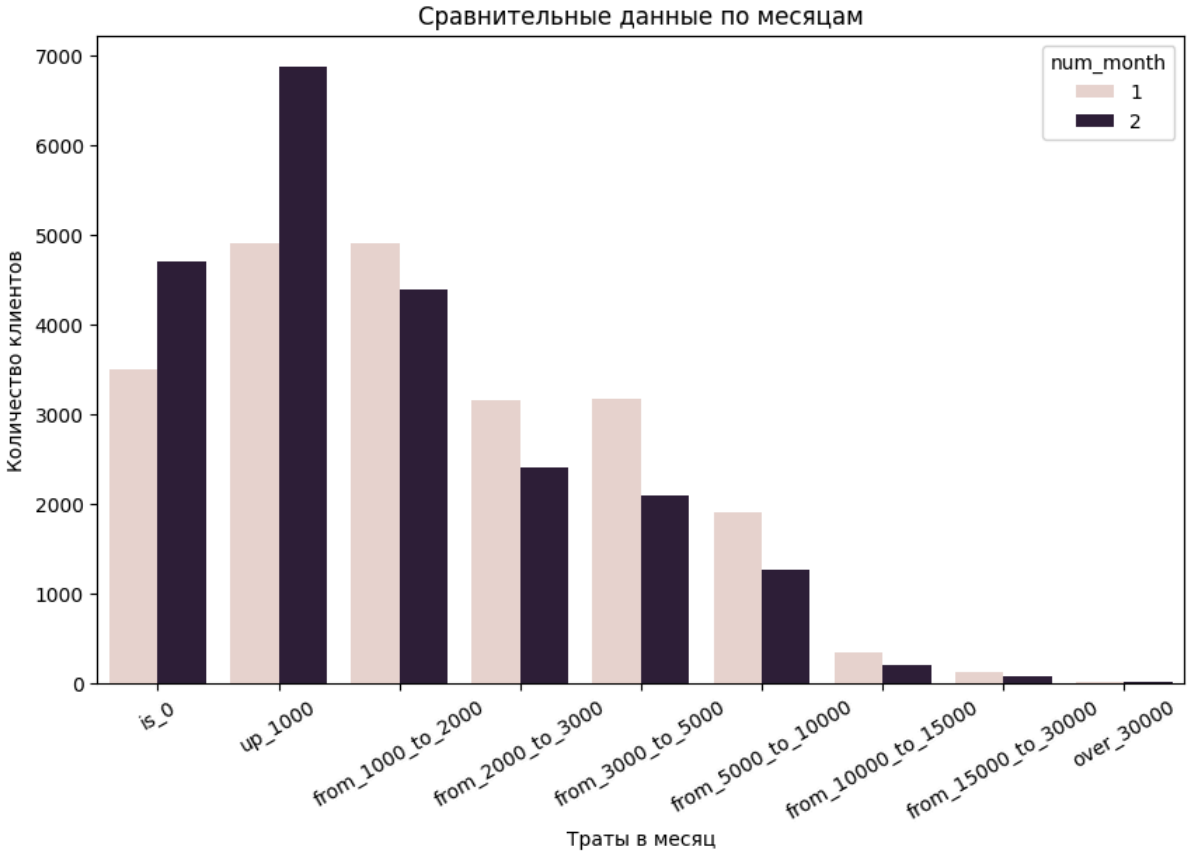


Так как в датасете с покупками ограниченный временной промежуток, можем на ограниченной выборке рассмотреть, как изменяется количество покупок в первый и второй месяц после приобретения карты лояльности.



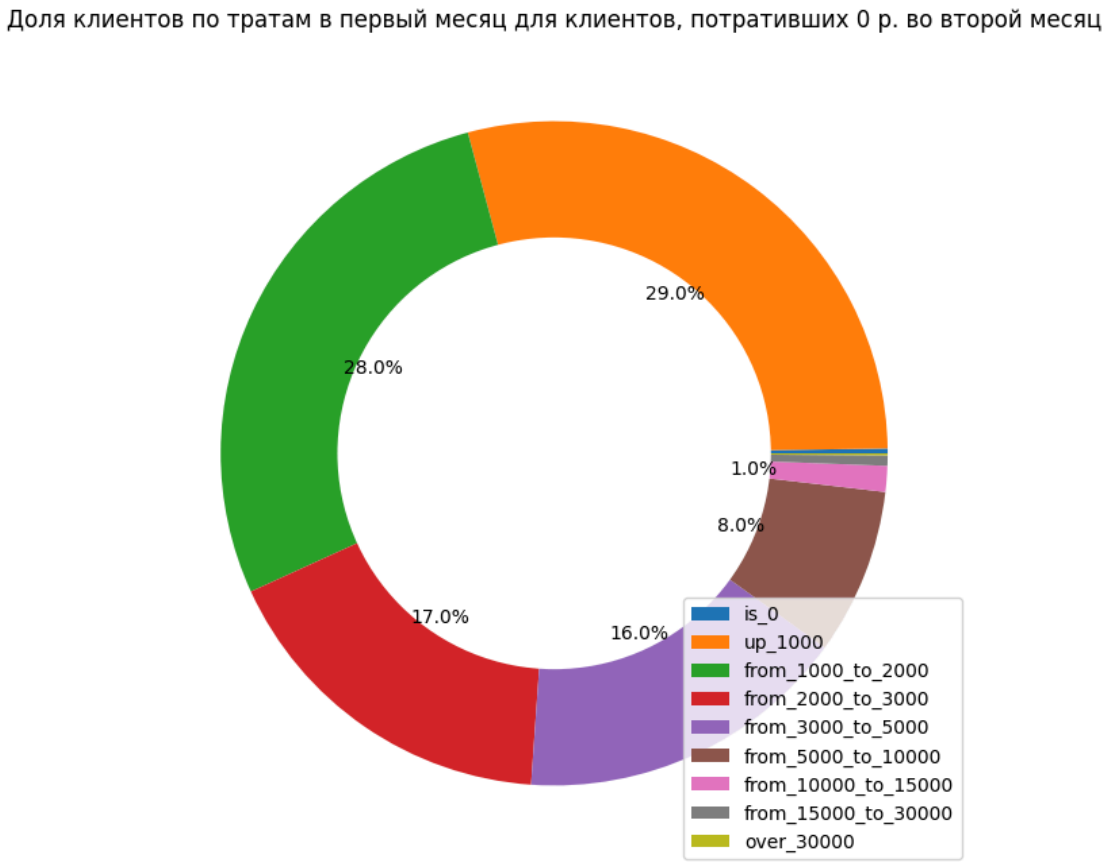


Как мы видим, отсутствует ярко выраженная зависимость от гендера распределений объемов затрат на покупки.



Видим, что

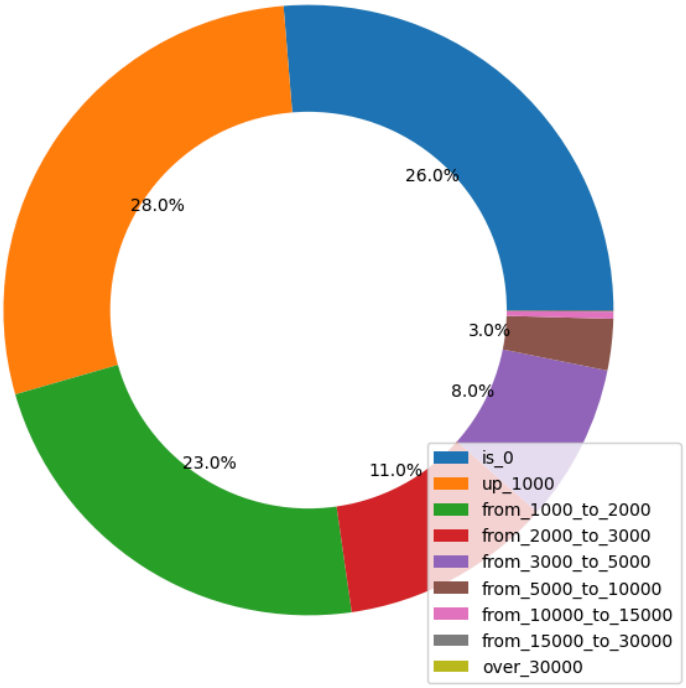
Теперь рассмотрим группу тех, кто тратит 0 р. в месяц, так как таких покупателей можем считать “потерянными”.



Удивительно, но доля клиентов, во второй месяц потративших 0 р., которые при этом потратили 0 р. в первый месяц меньше 1 процента. Более половины из клиентов, ставших “мертвыми” клиентами во второй месяц, имели траты до 2000 р.

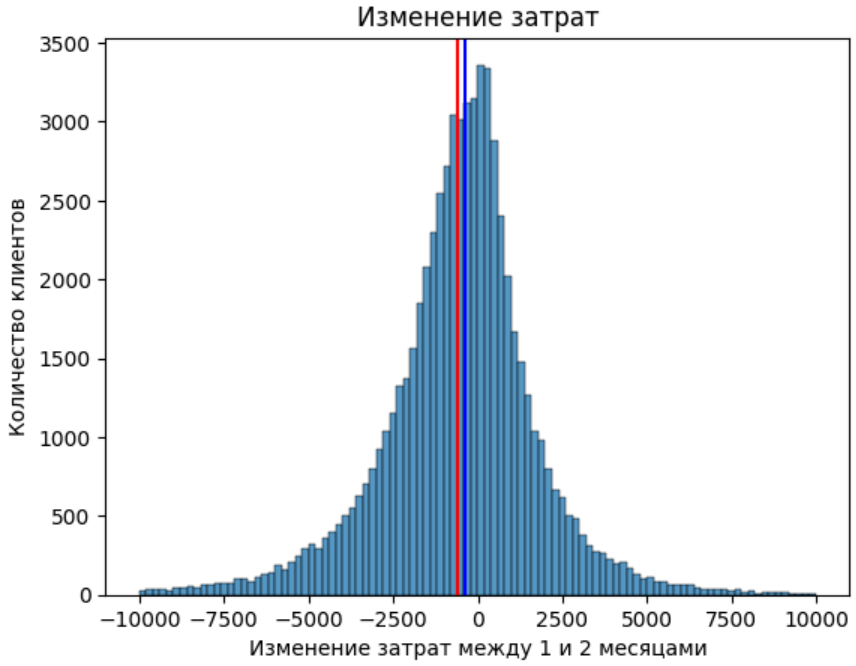
На следующей диаграмме можем увидеть, что клиенты, потратившие 0 р. во второй месяц, в основном, тратили 0 р. или до 2000 р.

Доля клиентов по тратам в первый месяц для клиентов, потративших до 1000 р. во второй месяц



На следующем графике мы видим, что в среднем на ~602 р. уменьшается затраты на покупки клиента во второй месяц, по сравнению с первым месяцем. Медианное значение = –397.0 р.

Рассмотрим поподробнее профиль изменения затрат.



Результаты применения OLS-регресии:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-86.3341	18.898	-4.568	0.000	-123.374	-49.294
1_month	0.2876	0.003	105.069	0.000	0.282	0.293
age	-0.4751	0.147	-3.224	0.001	-0.764	-0.186
count	67.7390	0.534	126.748	0.000	66.692	68.787
gender_M	-4.2053	23.289	-0.181	0.857	-49.852	41.441
gender_U	-16.6305	18.694	-0.890	0.374	-53.270	20.009

Видим, что гендер является незначимой переменной, исключим его.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-95.8930	14.044	-6.828	0.000	-123.420	-68.366
1_month	0.2875	0.003	105.174	0.000	0.282	0.293
age	-0.4762	0.147	-3.235	0.001	-0.765	-0.188
count	67.7617	0.534	126.961	0.000	66.716	68.808

Значения переменных:

- const – константа
- 1_month – затраты в первом месяце
- age – возраст клиента
- count – количество покупок
- gender_M – категориальная переменная (мужской пол)
- gender_U – категориальная переменная (неизвестный пол)

Как мы видим, 1_month, age, count устойчивы к удалению переменных, отвечающих за гендер, следовательно, можем считать их статистически значимыми.

Из данной регрессии можно сделать вывод, что возраст монотонно негативно влияет на новый объем затрат. Можем предложить, что существует точка минимума или максимума, при этом предположении введем переменную с квадратом возраста.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-84.1813	16.743	-5.028	0.000	-116.997	-51.365
1_month	0.2874	0.003	105.089	0.000	0.282	0.293
age	-0.7250	0.243	-2.981	0.003	-1.202	-0.248
count	67.7611	0.534	126.960	0.000	66.715	68.807
age**2	-6.02e-05	4.69e-05	-1.285	0.199	-0.000	3.16e-05

Видим, что гипотеза не подтверждается.

Рассмотрим регрессию на изменение 2_month–1_month.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-95.8930	14.044	-6.828	0.000	-123.420	-68.366
1_month	-0.7125	0.003	-260.594	0.000	-0.718	-0.707
age	-0.4762	0.147	-3.235	0.001	-0.765	-0.188
count	67.7617	0.534	126.961	0.000	66.716	68.808

Проинтерпретируем последний результат. Возраст негативно влияет на изменение объема затрат между месяцами, точно также как и объем покупок в первый месяц.

Понять данный результат достаточно просто: чем больше затрат в первый месяц, тем сложнее “поддержать тот же уровень” и во второй месяц. При этом, чем старше клиент, тем сильнее уменьшаются затраты.

Требуется поиск дополнительных параметров для большей статистической значимости константы и обучения нелинейных моделей.