

Проект: Анализ клиентской активности

Константинов Александр

Данный проект в сфере анализа данных выполнен на датасете "X5 Group", посвященном для решения совершенно другой задачи, однако мной использован для демонстрации технических и общих скиллов в области анализа данных.

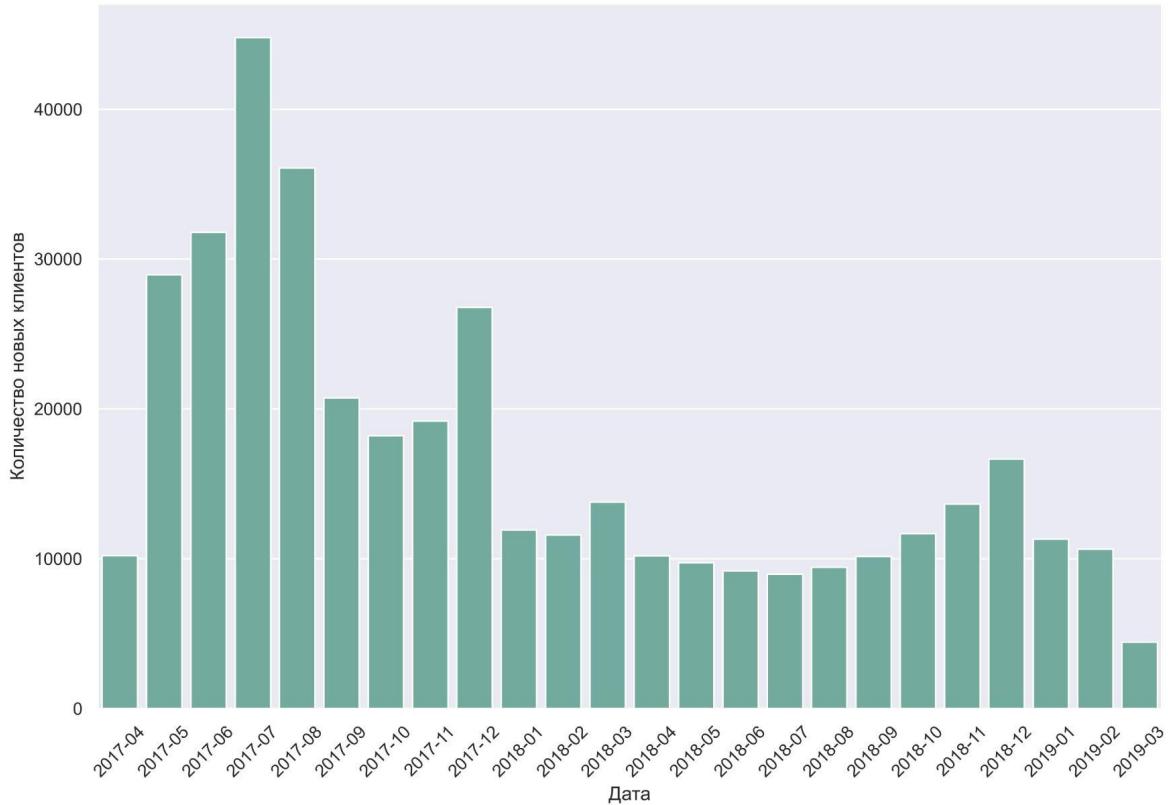
Датасет взят с Kaggle:

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

Общая информация о датасете

- Количество клиентов -- 400162
- Данные о клиентах, получивших карту лояльности с 04.04.2017 по 15.03.2019
- Гендерный состав: мужчины – 66 807, женщины – 147 649, не указан пол – 185 706
- Очевидно недостоверные данные о возрасте клиента – 1145. Например, значения -937, 119, 115 и т. д.
- Информация о транзакциях: количество транзакций – 8045201, количество строк с покупками – 45786568, период – с 21.11.2018 по 18.03.2019
- Принимая в расчет, что данные могут быть неполными, но выражающими общие тенденции генеральной совокупности, стоит понимать, что цены актуальны для 2018-2019 гг., поэтому месячные расходы клиентов в районе 3000-5000 рублей реалистичны в силу уровня цен, также некоторые транзакции могут не учитываться, так как клиенты могли не пользоваться картой лояльности при некоторых покупках.

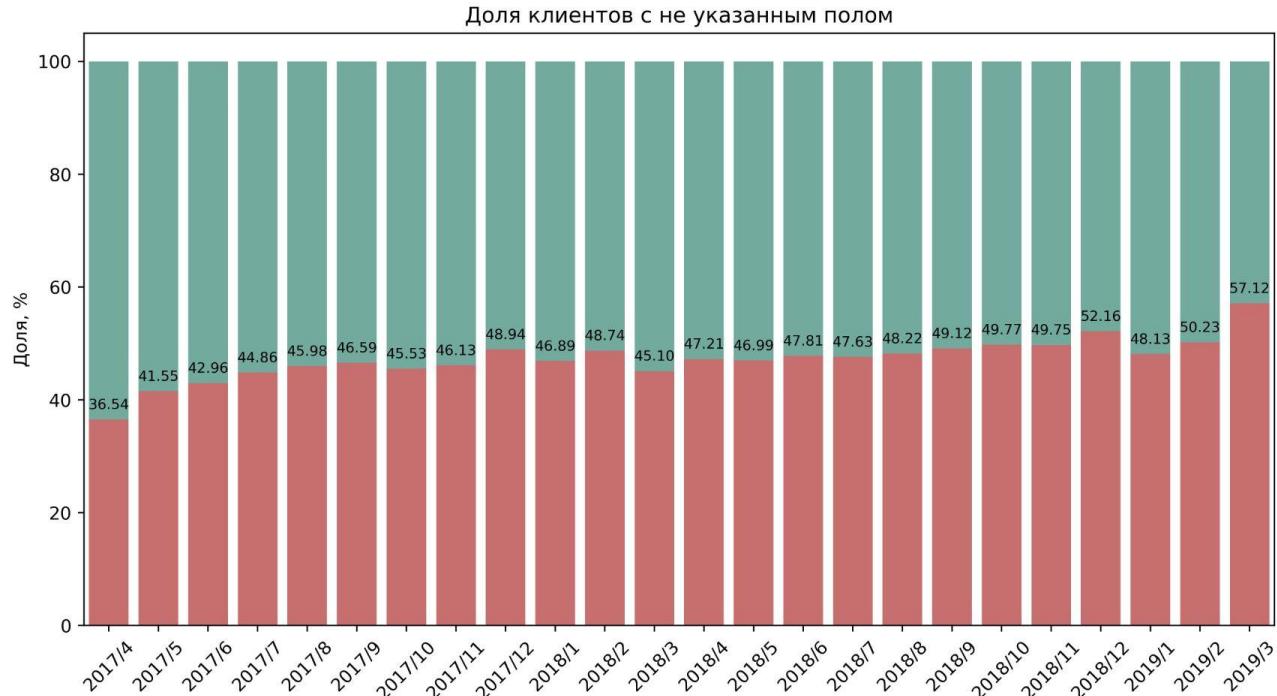
Динамика притока новых клиентов



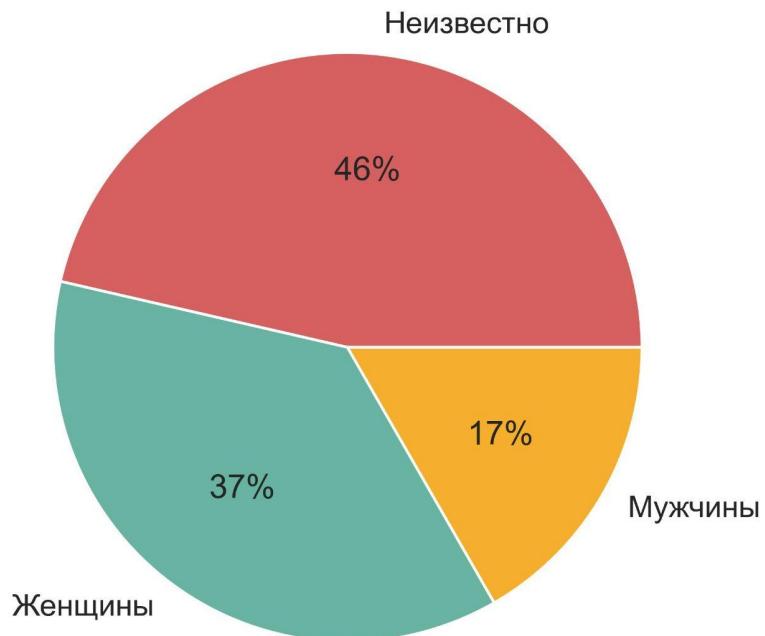
Как можно заметить,
с каждым годом
снижается приток
новых клиентов.

Проблема:

Рост количества клиентов, не
указавших пол.



Распределение клиентов по полу



В предположении о примерном равенстве доли мужчин и женщин среди клиентов можем предположить, что среди клиентов, не указавших гендер, больше мужчин, чем женщин.

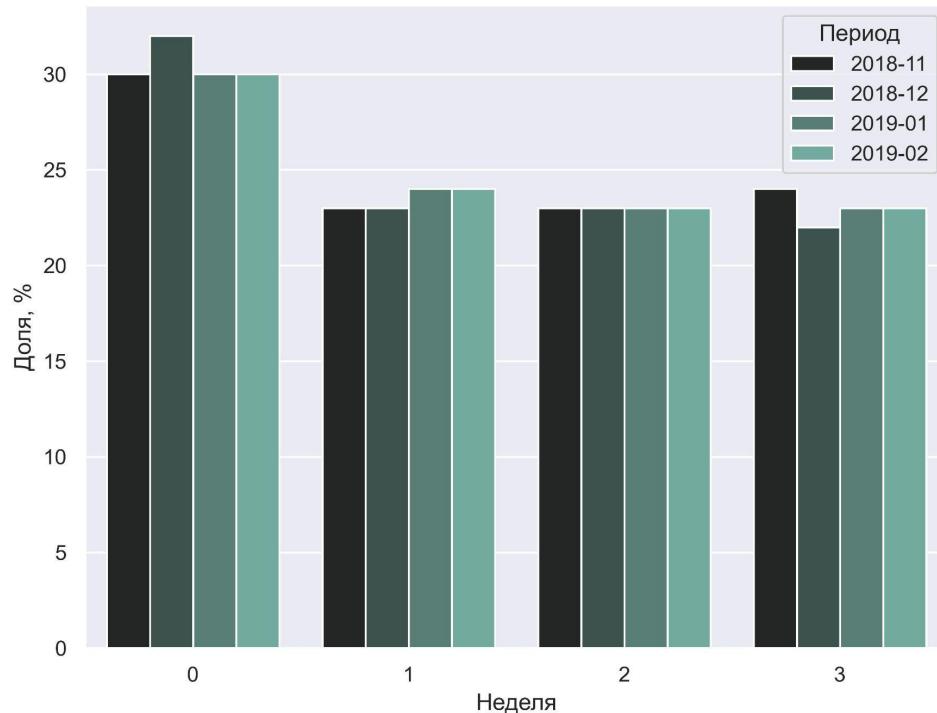
Как ведут себя новые клиенты?



Во-первых, мы можем наблюдать эффект новизны.

Средний расход за первый месяц *среди новых клиентов выше*, чем средние траты за выбранный месяц *среди старых клиентов*.

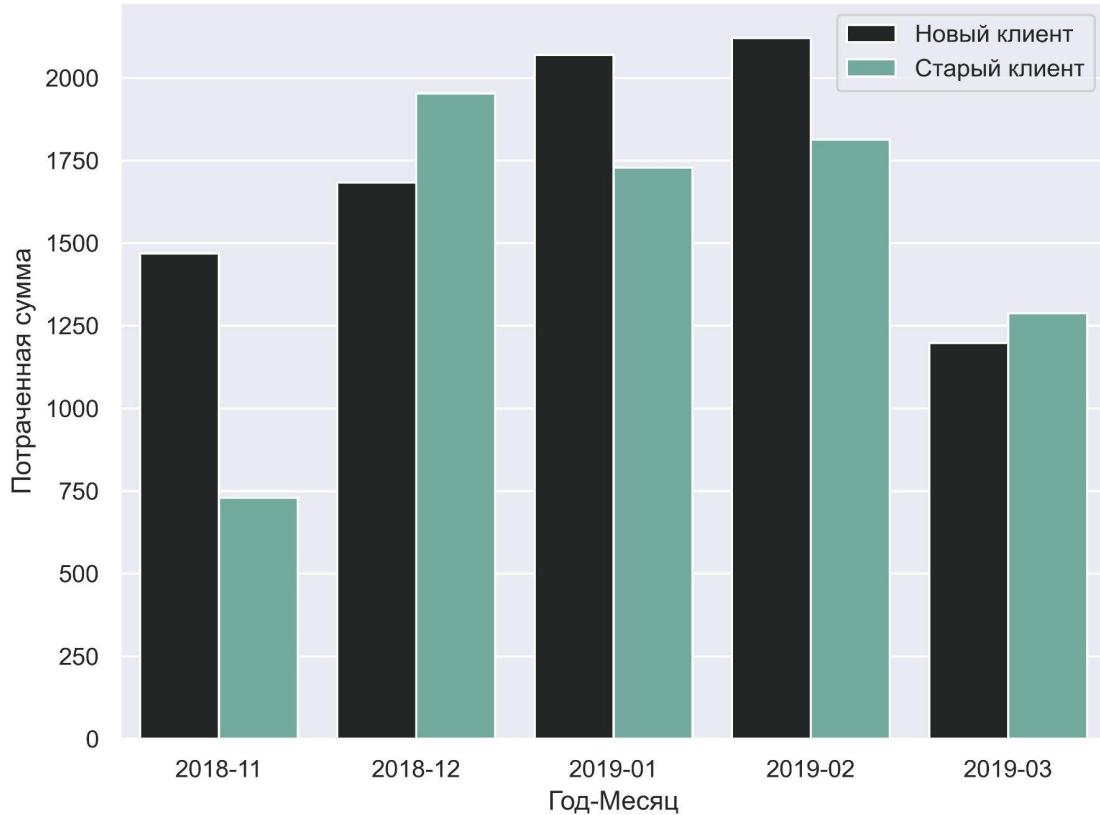
Как долго действует эффект новизны?



Оказывается, что примерно **1 неделю**.

Доля потраченной в 1 месяц суммы стабильно выше в 1 неделю (этот факт не зависит от месяца).

Рост потраченной суммы за первый месяц (без 1 недели)

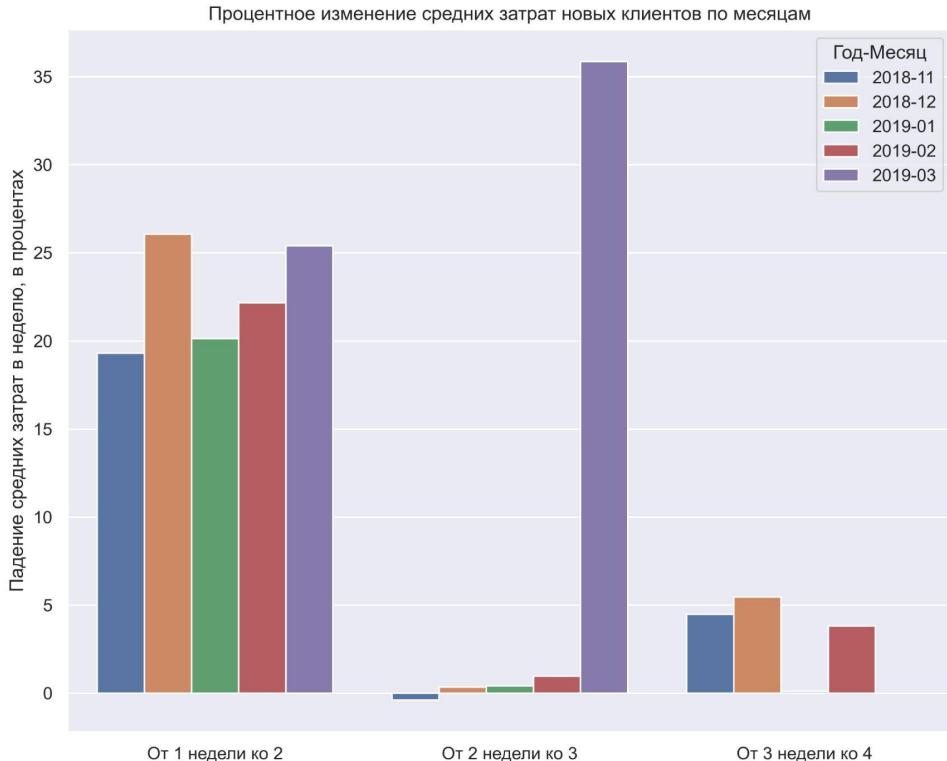


Расходы новых клиентов в первый месяц (за вычетом 1-ой недели) стремительно приближаются к средним значениям за 3 недели **для старых клиентов.**

Примечание:

Статистика за ноябрь 2018 и март 2019 неполная. Более того, для марта 2019 не существует новых клиентов, достигших 4 недели пользования картой лояльности (в силу ограничений датасета), для 3 недели данные неполные.

Поэтому мы можем видеть следующий эффект при визуализации:



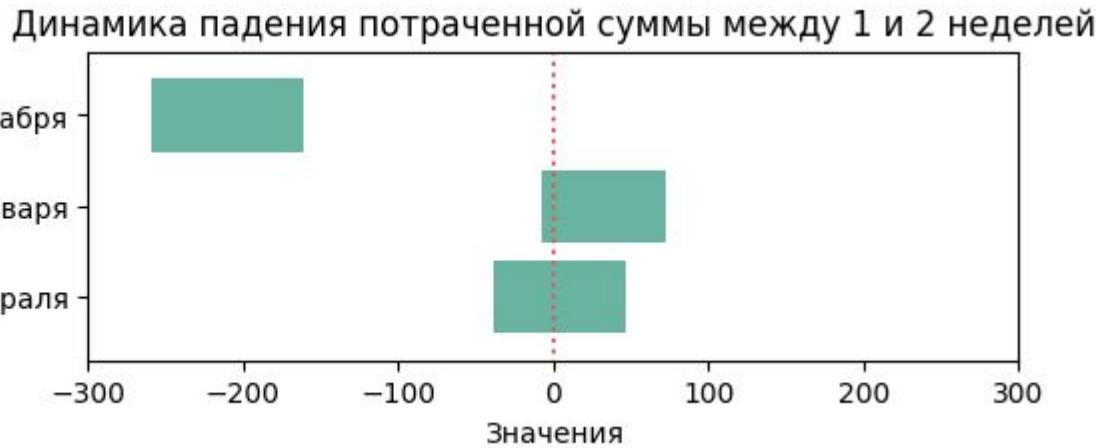
Во *вторую неделю*
клиенты тратят сумму
меньше от 20 до 25
процентов.

При переходе к *третей*
неделе и четвертой
неделе изменения
незначительны.

Падение на 35 процентов
для марта – следствие
неполноты данных.

Результаты статистических тестов

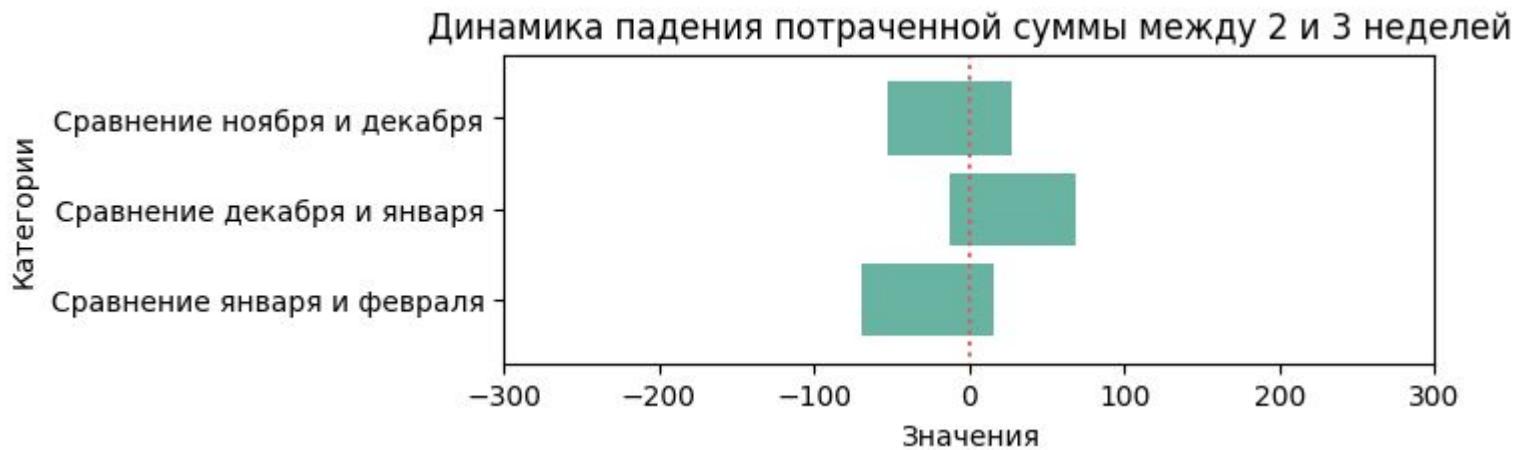
Категории



Как можно заметить, на 5-процентном уровне значимости между ноябрем и декабрем есть значительные изменения.

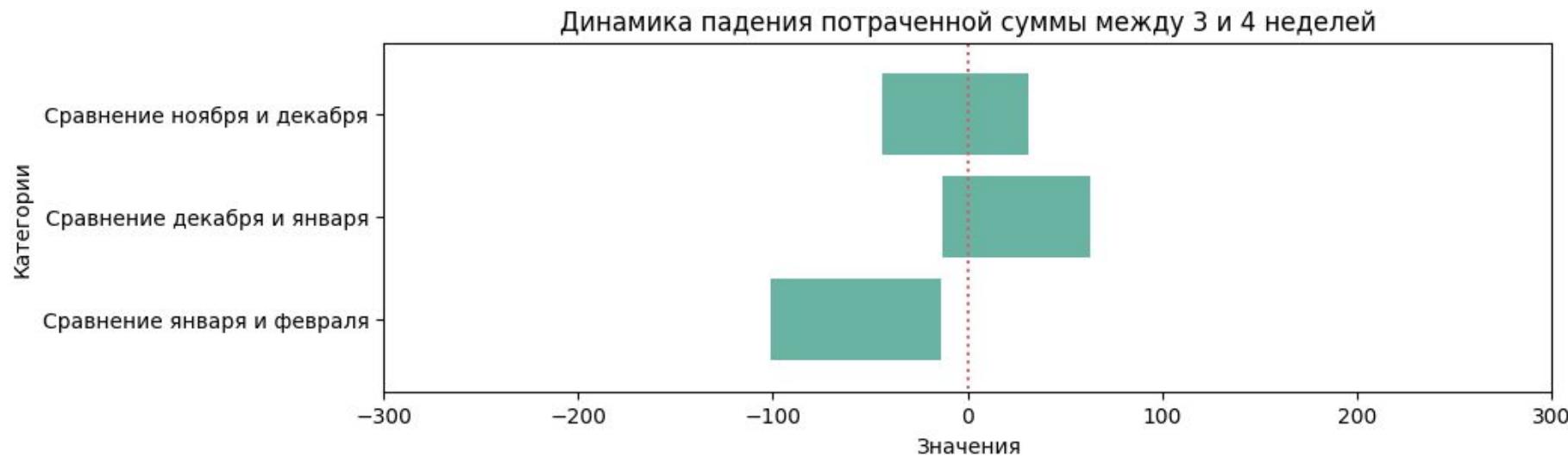
В декабре разрыв между первой и второй неделями трат значительно больше, чем в ноябре. Однако для января и февраля подобный вывод сделать не можем.

Ранее уже говорилось, что падение между 2 и 3 неделями близко к нулю. Соответственно, и разница средних не отличается от нуля статистически значимым образом.



Между 3 и 4 неделями клиентское поведение отличается, наблюдается некоторое падение в среднем. В районе 5 процентов, как мы видели ранее.

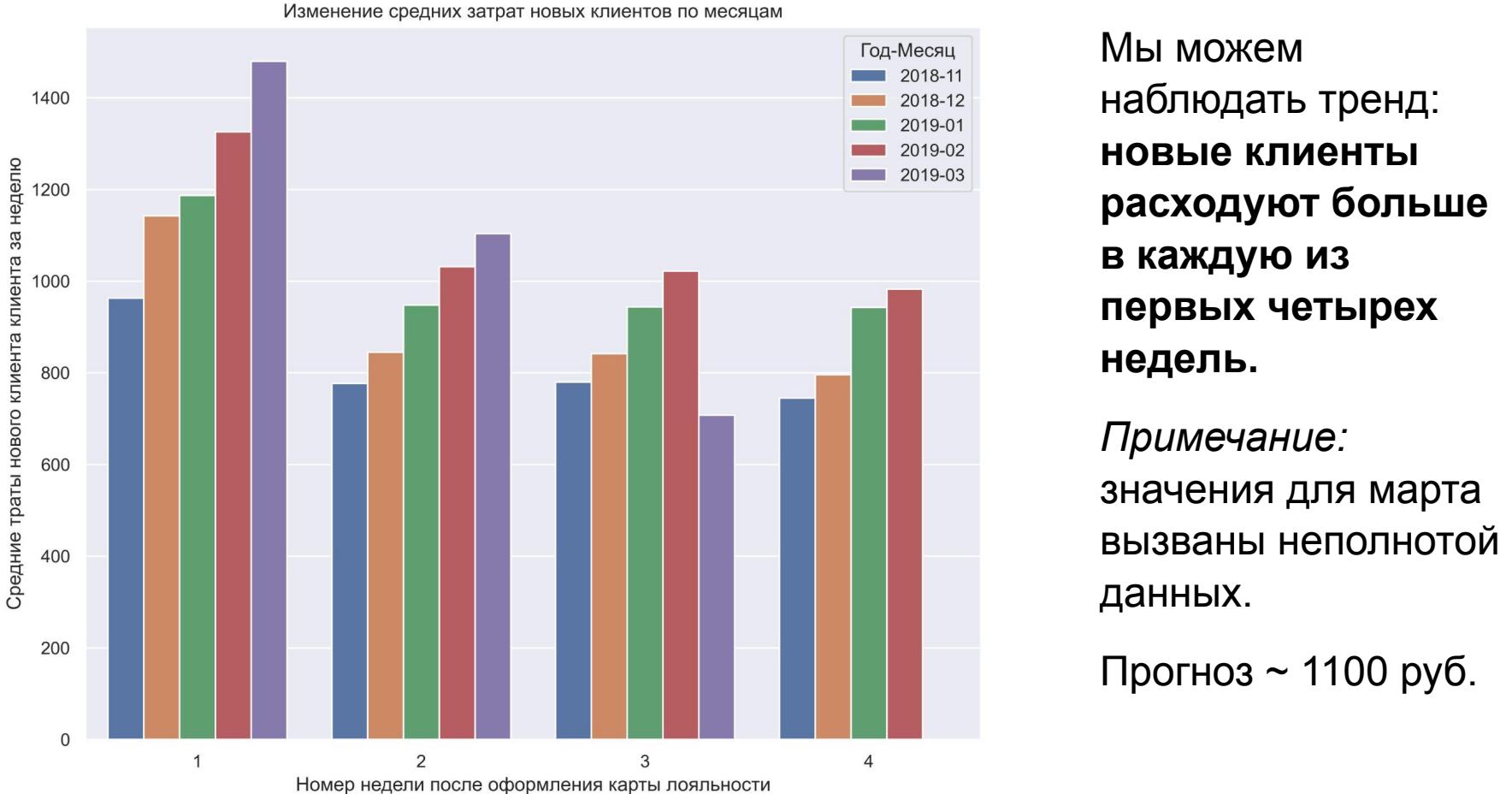
Так как для января это число равнялось 0, можно значимо сказать, что в феврале ситуация ухудшилась по сравнению с январем. Об улучшении в январе статистически значимо сказать не получается.



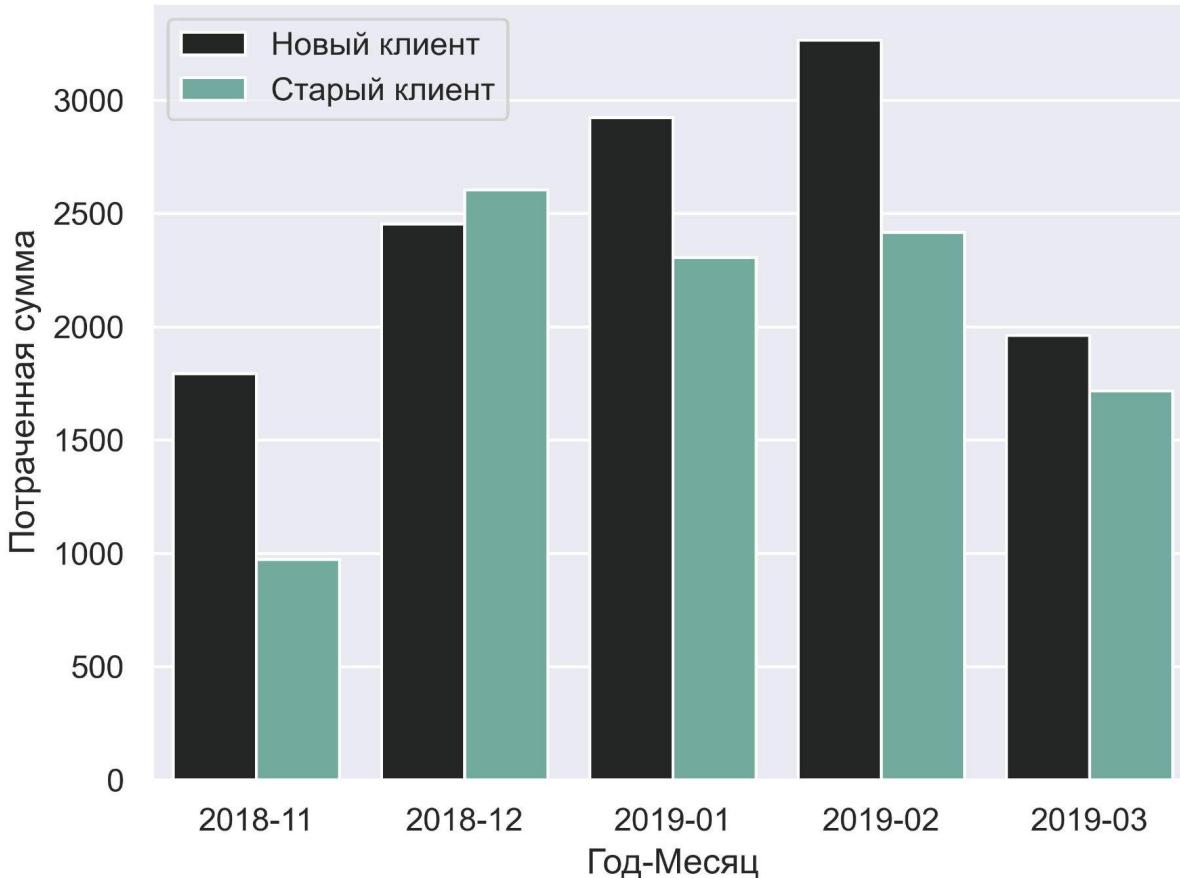
К обзору приложена полная картина. Изменение потраченной суммы между неделями подчиняется нормальному закону, что позволяло пользоваться t-тестом для сравнения средних. Так как дисперсии различны, был использован тест Уэлча.

Предыдущие слайды касались того, на сколько меняется потраченная новым клиентом сумма между неделями. Падения, оказываются, в большинстве своем, одинаковыми для каждого месяца.

Однако общий тренд – рост суммы, потраченной на каждой из первых 4 недель.



Рост потраченной суммы за первый месяц

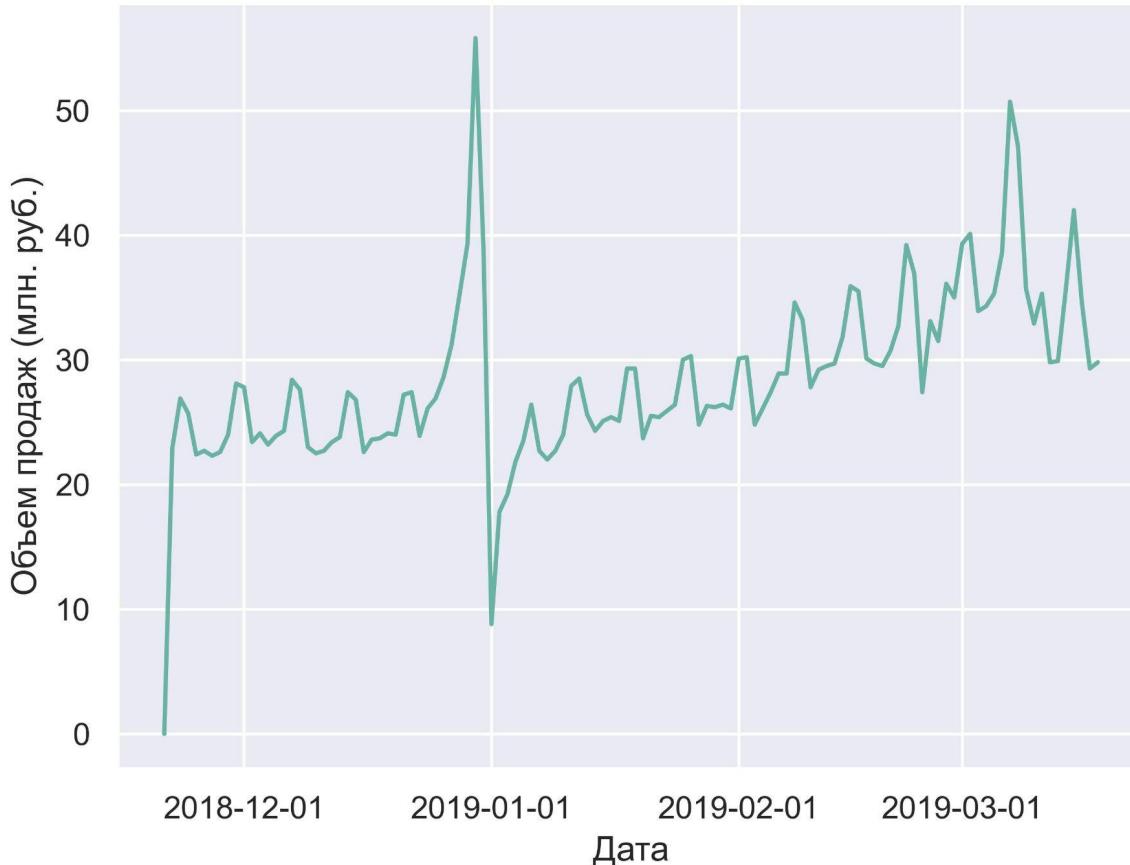


Как мы ранее видели: во все месяца, **кроме декабря**, средний расход нового клиента выше.

С чем это связано?

“Эффект 30 декабря”

Динамика продаж



Очевидно, что предновогоднее время – самое активное в плане покупок.

Пик продаж приходится на 30 декабря.

Невооруженным глазом эту точку видно на графике продаж.

Дата	Объем продаж
2018-12-30	55 759 508 руб.
2019-03-07	50 660 608 руб.
2019-03-08	47 104 855 руб.
2019-03-15	41 963 653 руб.
2019-03-02	40 138 267 руб.
2018-12-29	39 339 433 руб.
2019-03-01	39 335 485 руб.
2019-02-22	39 219 913 руб.
2018-12-31	39 006 354 руб.
2019-03-06	38 577 837 руб.

30 декабря
занимает **1**
место по
объемам
продаж.

Занимает **1**
место по
среднему
чеку.

Однако...

Дата	Средний чек
2018-12-30	616 руб.
2018-12-29	537 руб.
2019-03-08	520 руб.
2018-12-31	509 руб.
2019-03-07	491 руб.
2018-12-28	487 руб.
2019-02-23	477 руб.
2019-03-15	469 руб.
2019-03-02	467 руб.
2019-02-16	464 руб.

По количеству покупок занимает **2 место**, не сильно превосходя праздничное 8 марта и будние дни 1 марта, 11 марта и т. д.

При условии общего роста количества покупок мы видим от января к марту мы видим, что 30 декабря становится “аномалией временного ряда” в силу увеличения среднего чека.

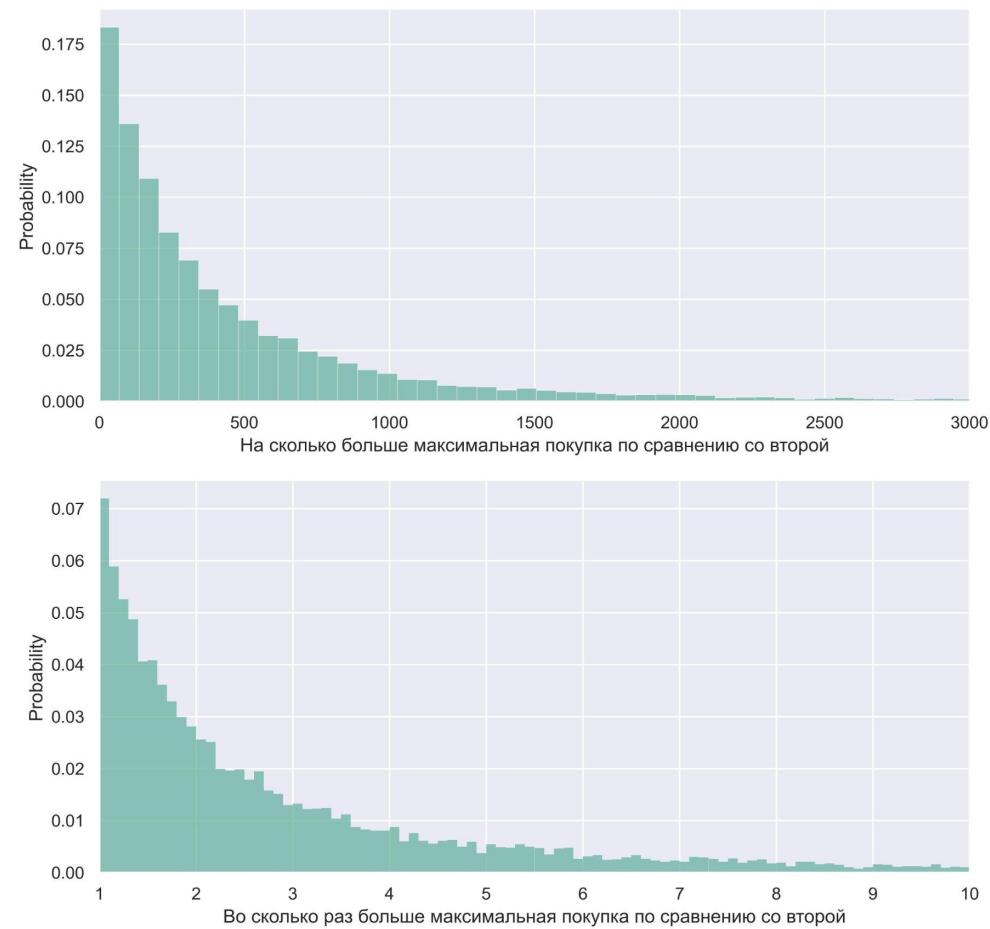
Дата	Количество покупок
2019-03-07	103 076
2018-12-30	90 469
2019-03-08	90 444
2019-03-01	89 465
2019-03-11	89 439
2019-03-15	89 300
2019-02-27	89 134
2019-03-04	87 192
2019-03-06	86 722
2019-03-05	85 919

Помимо того, что средний чек увеличивается, некоторые покупатели могут **повторно посещать магазин**: это не сильно влияет на рост количества покупок, однако **вторая по счету покупка равносильна по цене первой покупке.**

В итоге мы видим, что в течение всей предновогодней недели клиенты затрачивают больше денег.

Дата	Средняя трата на человека
2018-12-30	719 руб.
2018-12-29	613 руб.
2019-03-08	607 руб.
2018-12-31	596 руб.
2019-03-07	576 руб.
2019-02-23	554 руб.
2018-12-28	553 руб.
2019-03-02	538 руб.
2019-02-16	537 руб.
2019-03-15	531 руб.

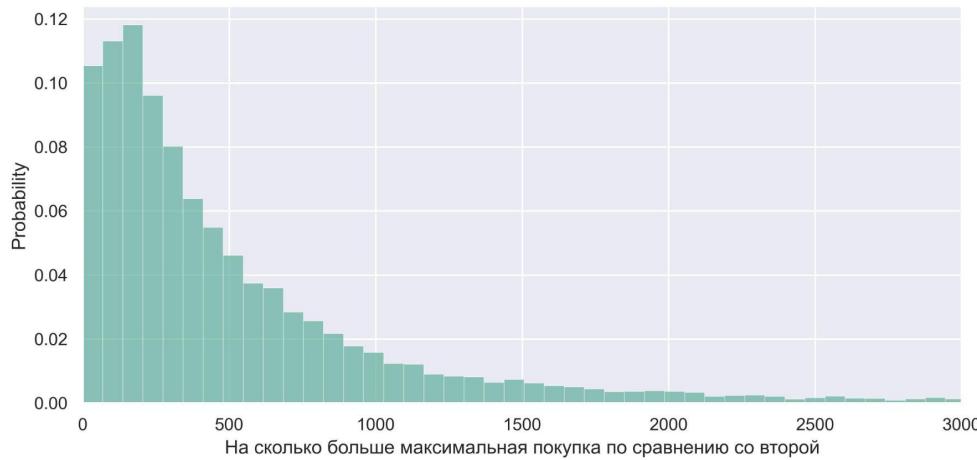
Сравнение максимальной покупки и второй по сумме



Как можно заметить
вероятность разницы
экспоненциально снижается с
высокой плотностью около 0
на первом графике и около 1
на втором графике.

Можем предположить, что
данная картина сильно зависит
от того, что многие клиенты
оформляют 2 и более мелких
покупок.

Сравнение максимальной покупки и второй по сумме (если максимальная покупка больше 200 р.)



Это предположение не подтверждается. Картина остается идентичной.

Таким образом, помимо роста среднего чека мы наблюдаем, что клиенты, совершившие более 1 покупки, совершают вторую покупку на сумму, не сильно отличающуюся от суммы максимальной покупки.

Вторым показателем, иллюстрирующим подобный факт, является **охват** уникальных клиентов.

По этому показателю 30 декабря занимает **7 место**.

Он оказывает менее значимое влияние.

Дата	Охват
2019-03-07	87 909
2019-03-11	80 023
2019-03-15	78 891
2019-03-01	78 657
2019-02-27	78 385
2019-03-08	77 586
2018-12-30	77 447
2019-03-04	76 821
2019-03-06	76 015
2019-02-25	75 969

Зависит ли прирост 30 декабря от возрастной группы?

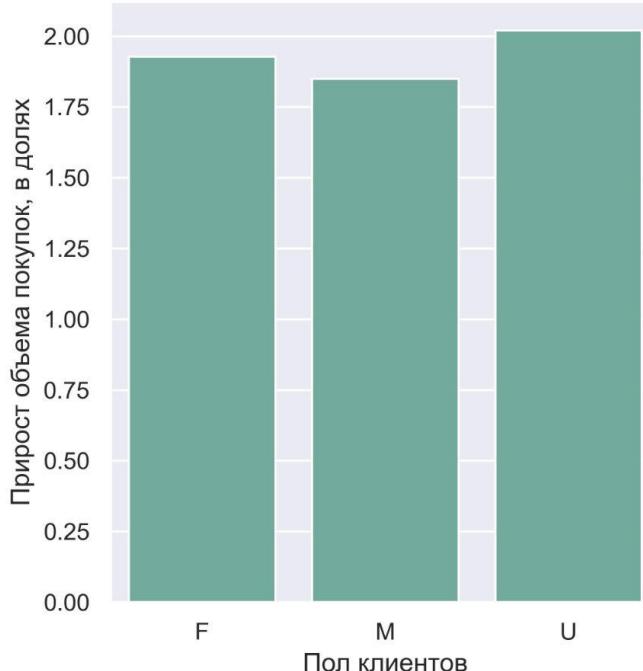
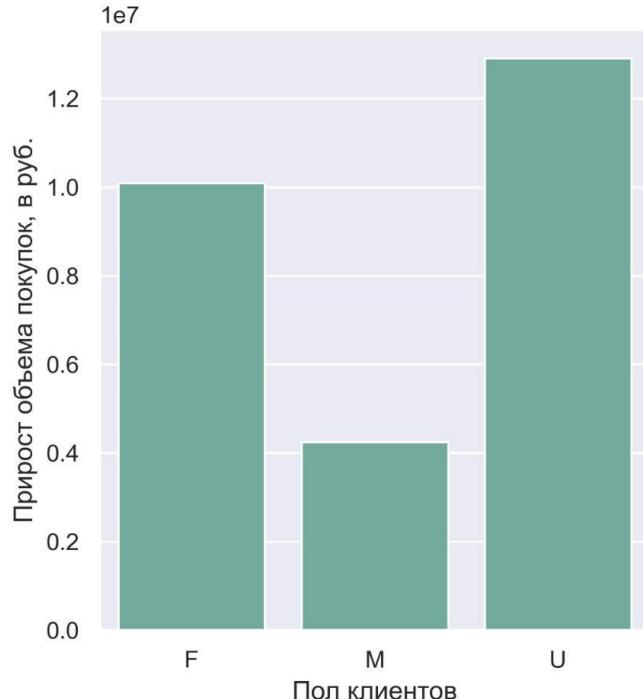
Нет, поскольку наблюдаемое неравенство прироста, исчисляемое в рублях вызвано распределением клиентом по возрасту.

В действительности, расходы клиента 30 декабря в среднем **в 2 раза выше**, чем в обычный день. Показатель незначительно колеблется в зависимости от возраста.



Зависит ли прирост 30 декабря от пола клиента?

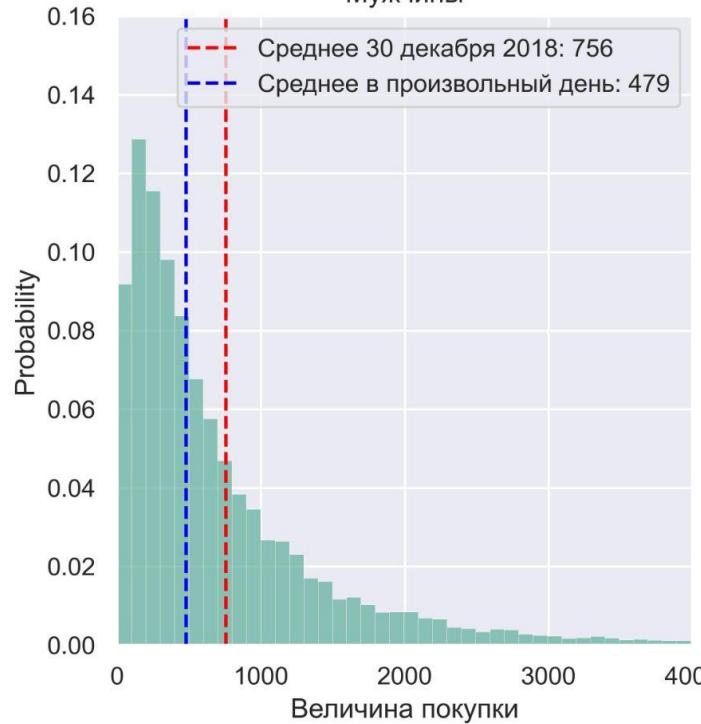
Прирост 30 декабря в зависимости от пола клиента



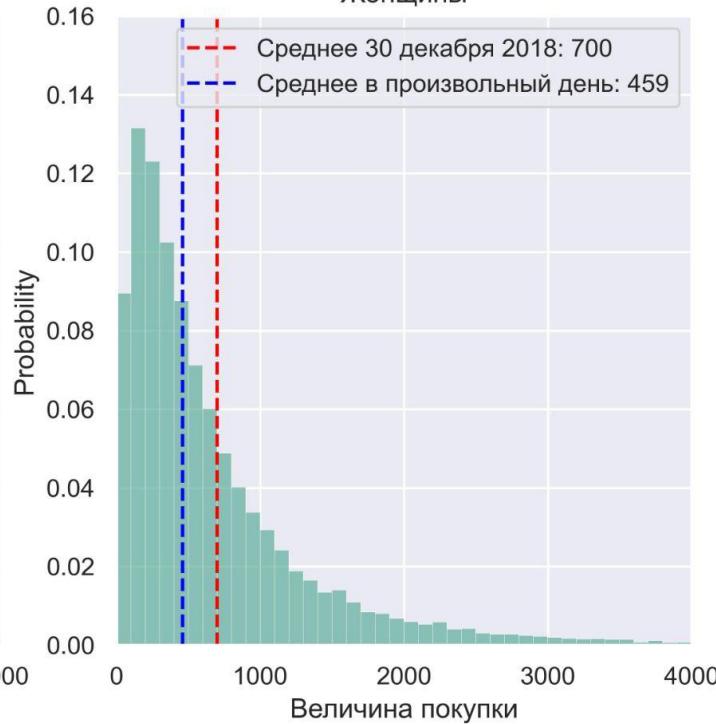
Мы видим, что показатели общего прироста затрат выросли примерно в равной степени для всех гендерных групп.

Картина значений в рублях – следствие гендерного дисбаланса клиентской базы.

Мужчины



Женщины



Прирост во
многом
происходит за
счет роста
среднего. Но мы
не видим
увеличения в 2
раза.

При построении бутстрэповского интервала для среднего не попадают значения, которые больше обычного в 2 раза.



Причина: плотность потраченной суммы около среднего ~ 750 р. Покупатели реже совершают мелкие покупки, чем в обычный день.

