Безальтернативная гипотеза

Чаще платишь, дальше едешь



Структура данных

- **1 017 640 записей**

Данные о клиентах банка и их тратах



Количественные переменные

- Расходы и количество заказов в “Городе”
- Расходы по категориям
- Возраст
- Количество детей
- Вероятности наличия машины, различных хобби

Структура данных

Бинарные переменные

- Пол
- Наличие питомца
- Занятия спортом
- Клиент города

Категориальные переменные

- Работа
- Регион



Структура данных

Город - сервис ежедневных покупок от Т-Банка

В датасете представлены сервисы:

- **Топливо**
- **Афиша**
- **Продукты**

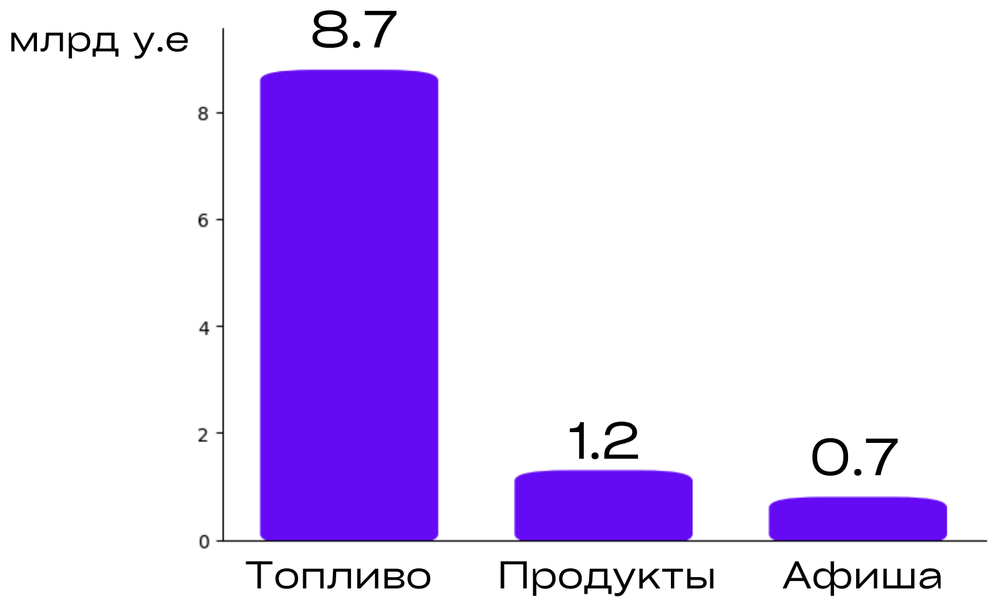


Структура данных

Процентные доли клиентов сервисов в датасете



Оборот в сервисах “Города”



Структура данных

Мы решили сконцентрироваться на **Топливе**

- Оборот **8.7 млрд** у.е
- **379215** пользователей

37% от всех

Кэшбэк

7% первый заказ → 1% последующие





Структура данных

01.01.2024 — 22.11.2024

Статистика трат по категориям
за последний квартал

- Статистика трат в Городе собрана **за последний год**



Оставляем клиентов топлива

Убираем

- Строки-дубликаты*
- Тех, кто не пользовался топливом
- Нет данных о транзакциях
- Тех, у кого нет машины
- Людей младше 18 лет или старше 100 лет

Осталось

351 851 записей

*см. приложение E

Предварительный анализ

Повышенный кэшбэк

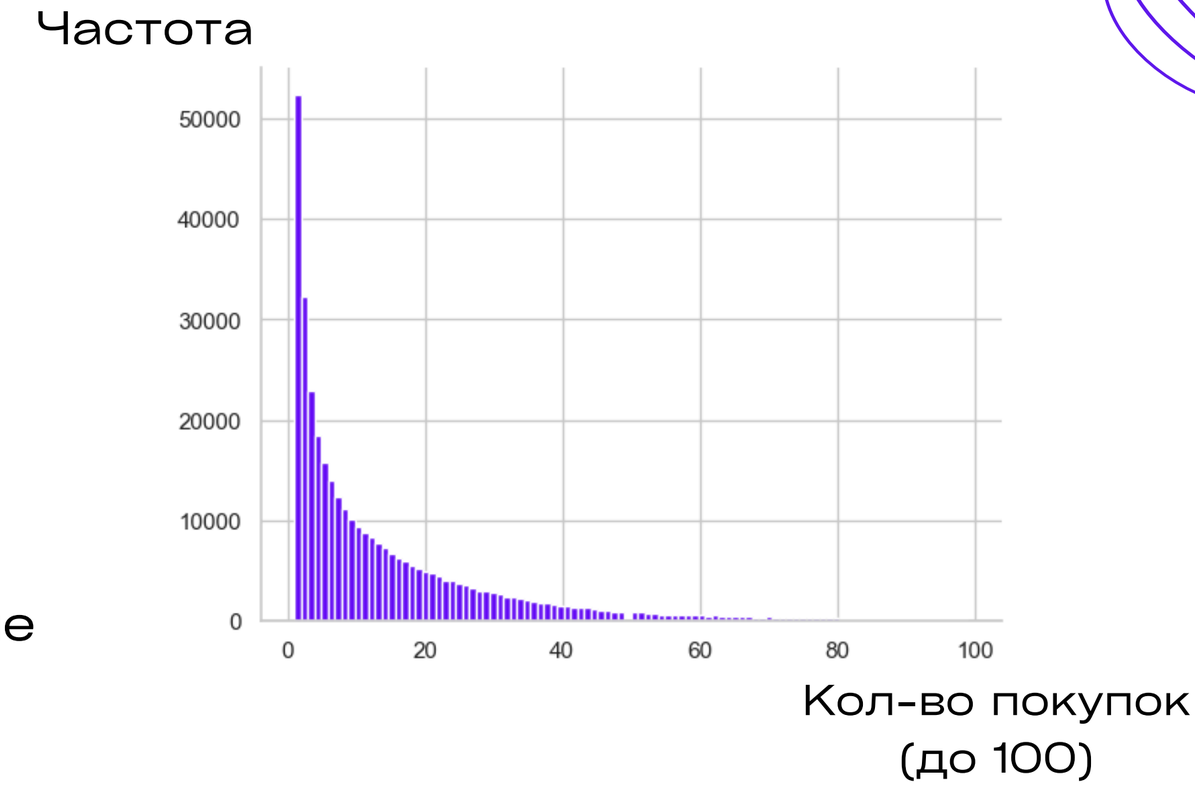
- первый заказ
- спецпредложения

Кэшбэк

7% → **1%**

первый заказ → последующие

Распределение **количества покупок** в топливе



Предварительный анализ

Медианная частота заказов = 1 заказ в **27.6 дней**

- только 1 покупка
- с нее прошло время, которое **меньше средней частоты заказов**

Нельзя говорить о том, что человек **не сделает второй заказ**



Удаляем их

Предварительный анализ

Новые переменные

second_flg

Сделал ли человек второй заказ после первого

transactions_cnt

Количество транзакций за последний квартал

economic_region

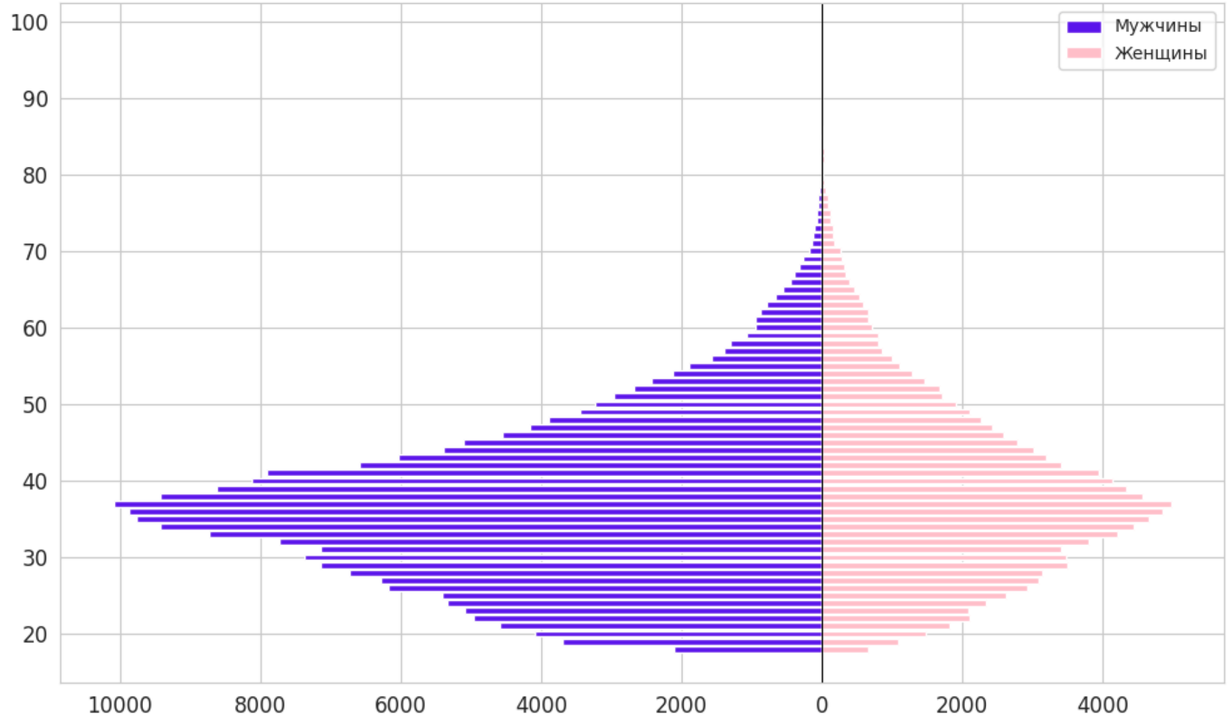
Экономический регион пользователя

(+ выделили Москву и Санкт-Петербург)

Предварительный анализ

Половозрастная пирамида

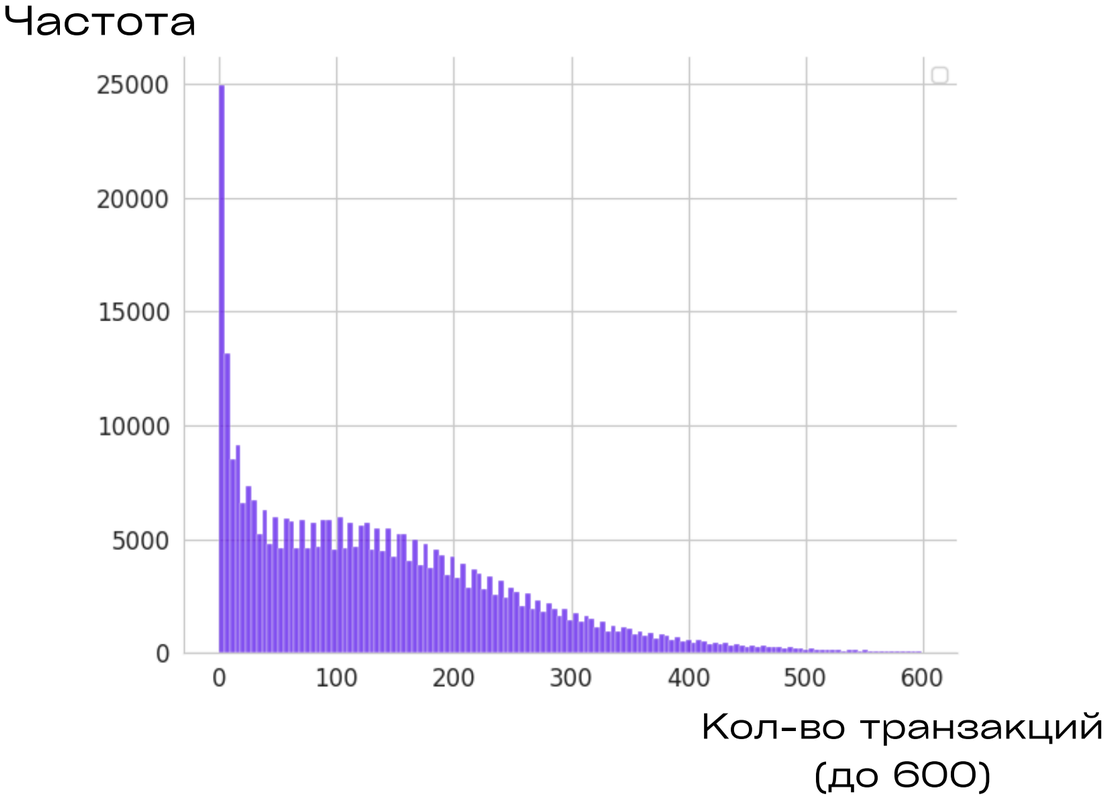
Возраст



Медиана возраста: **36**

Предварительный анализ

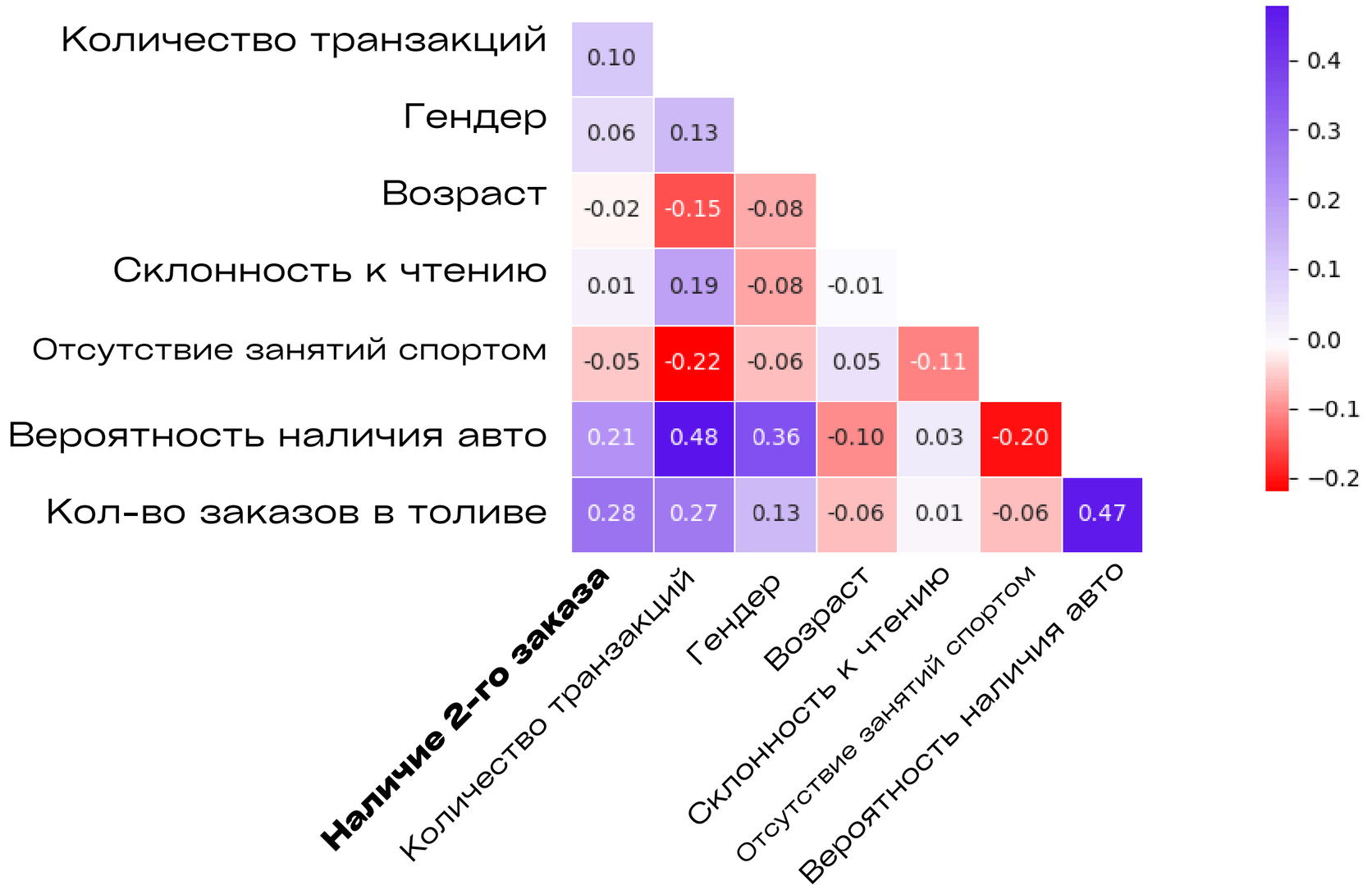
Распределение
количества транзакций
клиентов



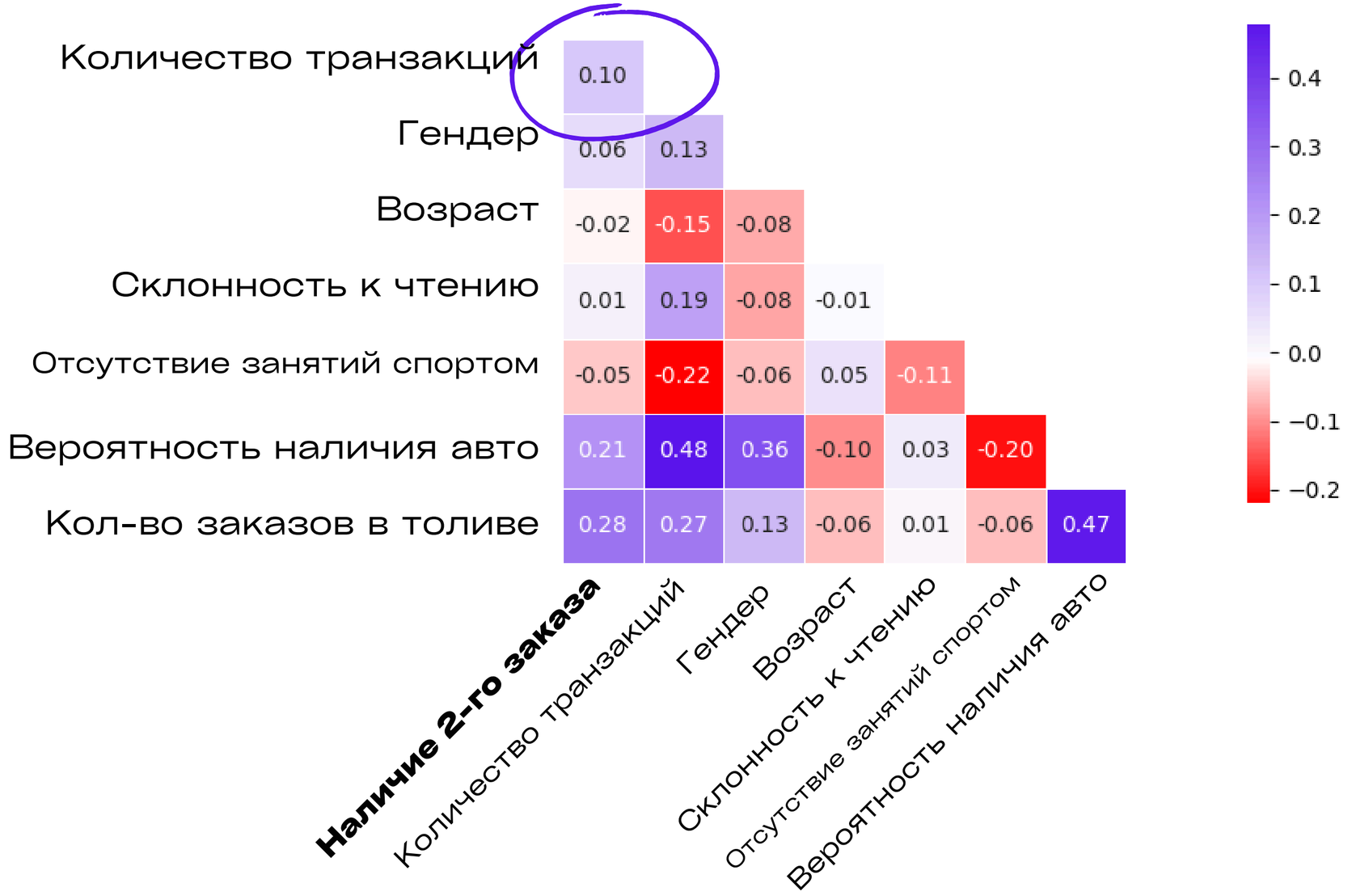
Сколько заказали
второй раз
(second_flag)



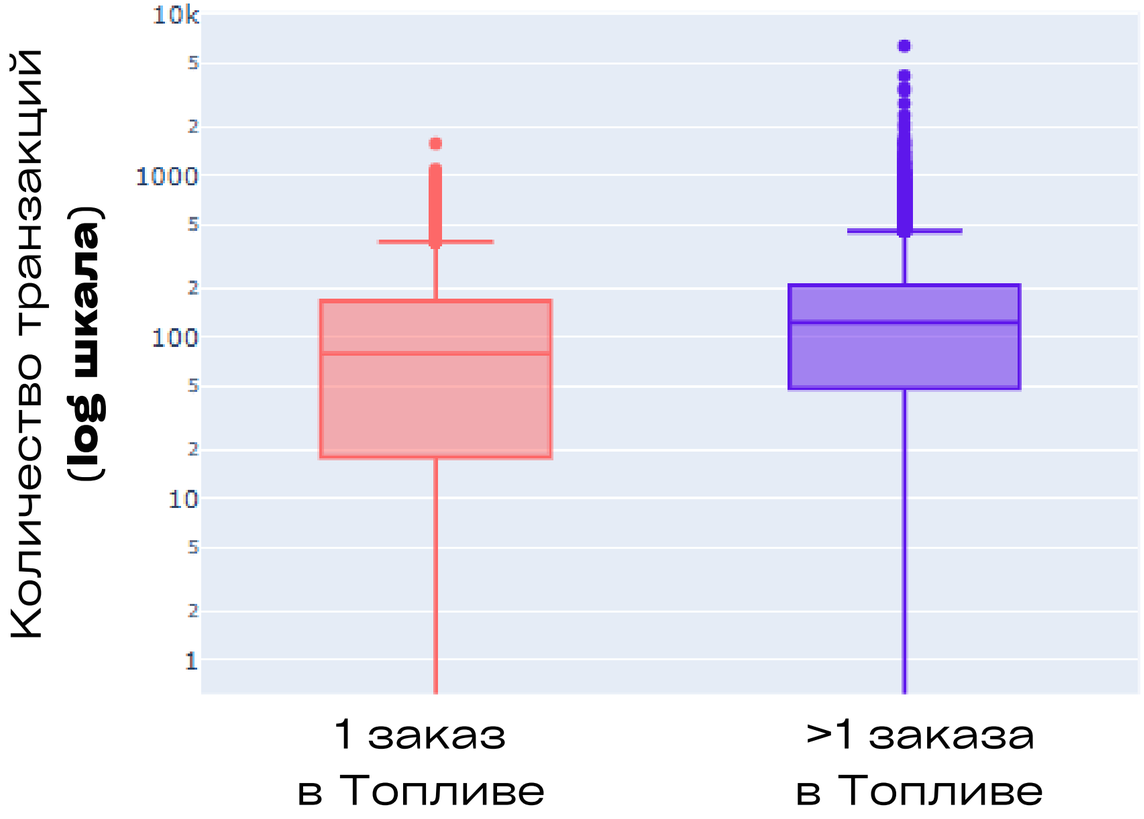
Предварительный анализ



Предварительный анализ



Предварительный анализ



Среднее количество транзакций у людей с **1 заказом в топливе**

109

Среднее количество транзакций у людей с **>1 заказа в топливе**

144

Как активность влияет на то, сделает ли человек второй заказ после первого?

Гипотеза

Активность* использования человеком карт банка положительно влияет на конверсию во второй заказ

*Активность – количество транзакций за определенный период (квартал)

Механизм

Клиент чаще покупает

Больше клиентский опыт = больше доверия

Больше вовлечённость в использование экосистемы компании

Остается после первого заказа

Математическое подтверждение

H0. Количество транзакций **не связано** с тем, что клиент сделает второй заказ

H1. Количество транзакций **связано** с тем, что клиент сделает второй заказ

$\alpha = 0.05$



Математическое подтверждение

Логит модель

Для проверки гипотезы мы решили выбрать **логистическую регрессию** → так как **таргет бинарный**

$$\hat{y} = \frac{1}{1 + e^{-(\beta + a \cdot x)}}$$

second_flg

0 - не сделал второй заказ
1 - сделал второй заказ

Математическое подтверждение

Факторы

- **Количество транзакций**
- Пол *
- Возраст

Таргет

Сделал ли человек второй заказ

Добавили **контрольные переменные** во избежание **смещения оценок** в результате пропуска существенной переменной

*см. приложение А

Математическое подтверждение

Переменная	Coefficient	P-value
Количество транзакций	0.0027	0.000
Возраст	0.0309	0.000
Возраст^2	-0.0004	0.000
Пол	0.2717	0.000

Все p-value **< 0.05**

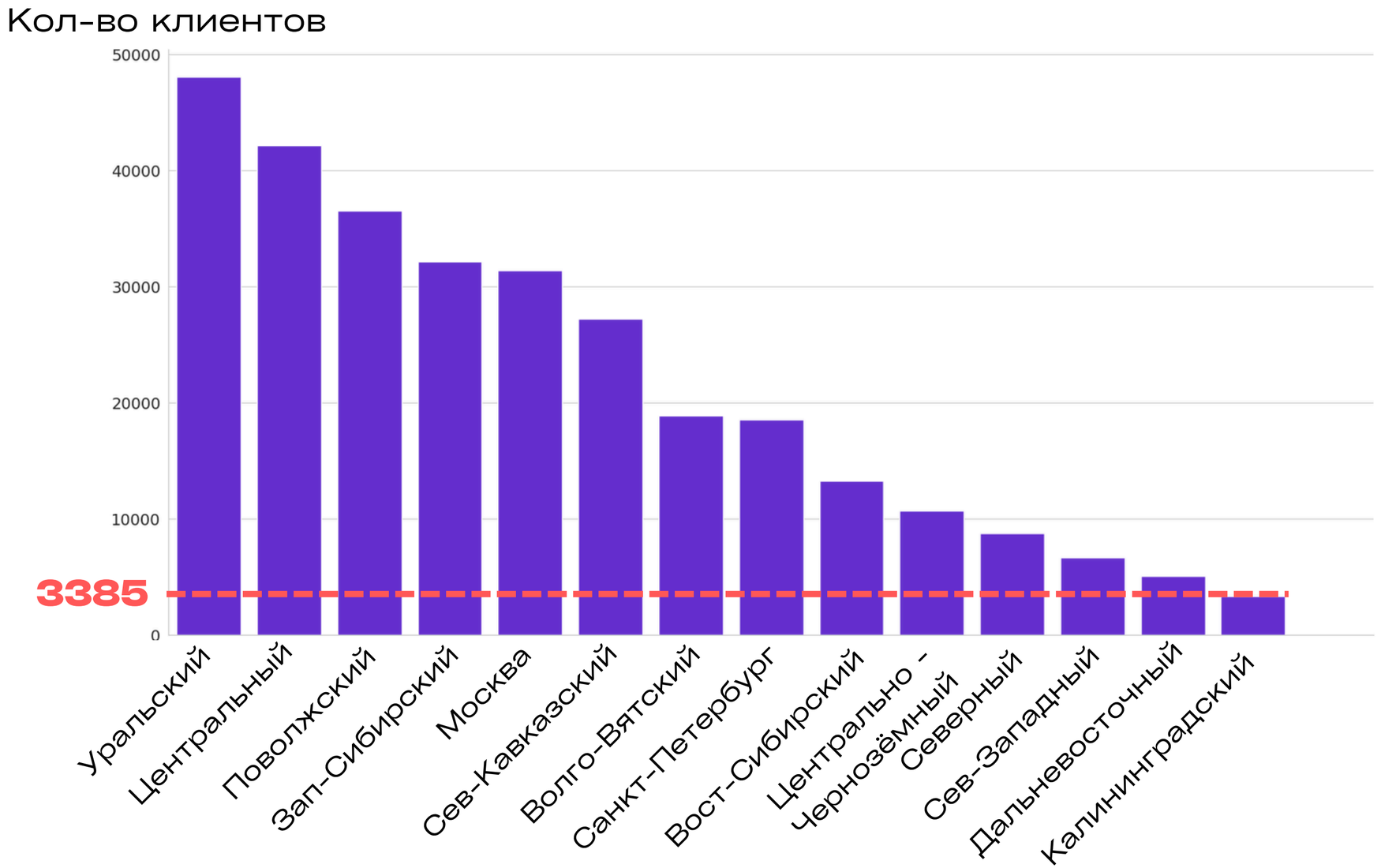
Принимаем **альтернативную гипотезу**

Проверка устойчивости

- По экономическим районам



Распределение по районам



Устойчивость по районам

Район	Коэффициент	P-value
Северо-Западный	0.0022	0.000
Уральский	0.0029	0.000
Западно-Сибирский	0.0026	0.000
Северо-Кавказский	0.0027	0.000
Центрально-Черноземный	0.0029	0.000
Поволжский	0.003	0.000
Москва	0.0019	0.000

Район	Коэффициент	P-value
Волго-Вятский	0.0028	0.000
Центральный	0.0025	0.000
Восточно-Сибирский	0.0039	0.000
Санкт-Петербург	0.0022	0.000
Калининградский	0.0035	0.000
Северный	0.0028	0.000
Дальневосточный	0.0019	0.000

Уровень значимости = 0.05. p-value везде **< 0.05**,
альтернативная гипотеза подтверждается

Интерпретация

- Между активностью пользователя и тем, выберет ли он воспользоваться Топливом второй раз, существует **статистически значимая положительная взаимосвязь**

- Модель устойчива и работает во всех экономических районах, т.к. при разделении p-value меньше уровня значимости и знак коэффициента сохраняется

Чаще покупает → **Остается в топливе**

Альтернативный механизм

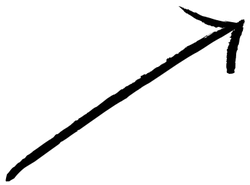
Клиент чаще покупает



Чаще заходит в приложение



Чаще вспоминает про сервис Топливо



Остается после первого заказа

Практическая польза

Исследование связи активности пользователя и конверсии во второй заказ в сервисе Город



- Банк сможет привлечь больше активных пользователей в сервис Город

Policy implication

Более активным выше стартовый кэшбэк, чтобы они начали пользоваться

Менее активным ниже стартовый кэшбэк, чтобы меньше терять

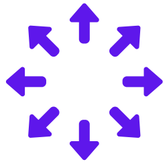
Оптимизация трат

Ограничения





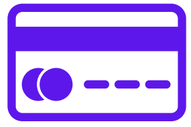
Не знаем о активности пользователя в течение всего года (только квартал)



Нельзя распространить исследование на другие подобные сервисы



Не знаем, когда клиент начал пользоваться карточкой банка



Не знаем, какая карточка у клиента*

*см. приложение D

Перспективы

Проверить пункты механизма

- Действительно ли **чаще покупающие чаще заходят в сервис через рекламу в приложении**

(Собрать данные о переходе в Топливо из рекламы в приложении)

- Действительно ли **чаще покупающие люди больше доверяют банку**

(Провести опрос о доверии банку)

Команда

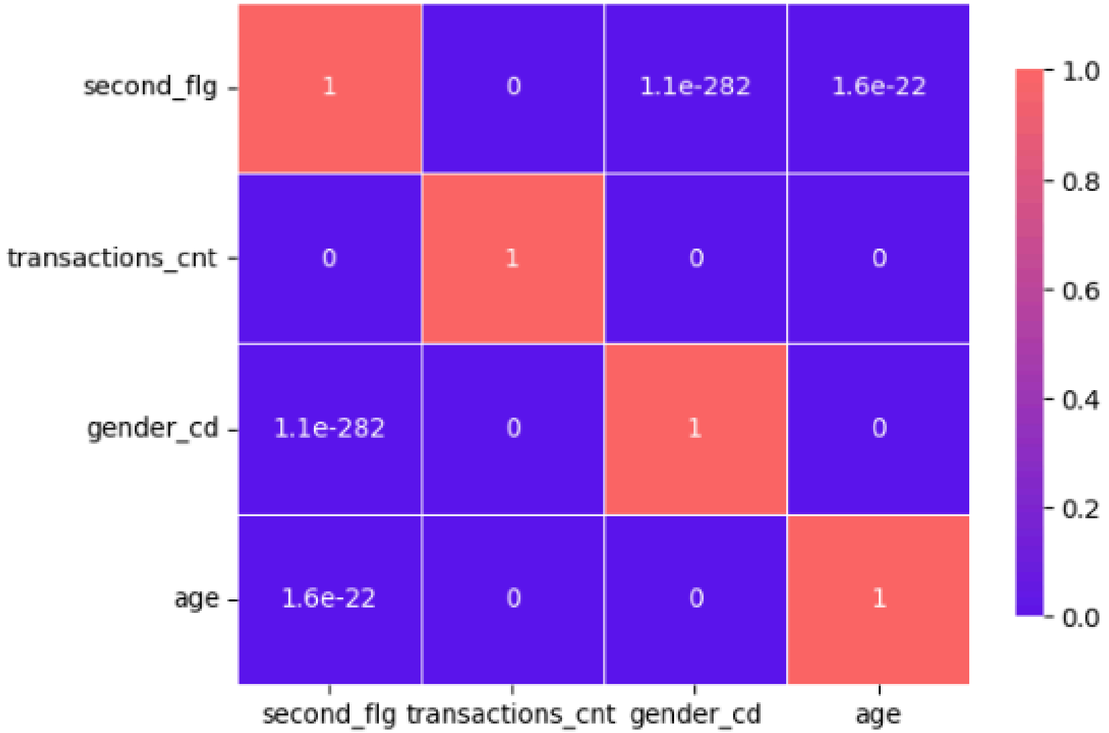
Безальтернативная гипотеза

СХ	АГ	ЛД	СК	ММ
Хватов Сергей	Галеев Артур	Дуванов Лука	Кулибаба Степан	Мерзликин Матвей

GitHub
с кодом

Приложения

Почему именно эти контрольные переменные?



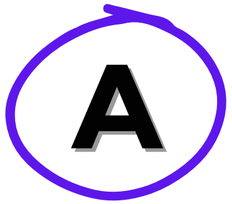
- **Статистически значимая корелляция с second_flg**

- Различия в экономическом поведении **по полам**

https://cyberleninka.ru/article/n/gendernye-razlichiya-v-potrebitelskom-povedenii-i-kommunikatsiyah/viewer

- Различия в экономическом поведении **по возрастам**

https://cyberleninka.ru/article/n/kriterii-i-priznaki-segmentatsii-potrebiteley-bankovskih-uslug/viewer

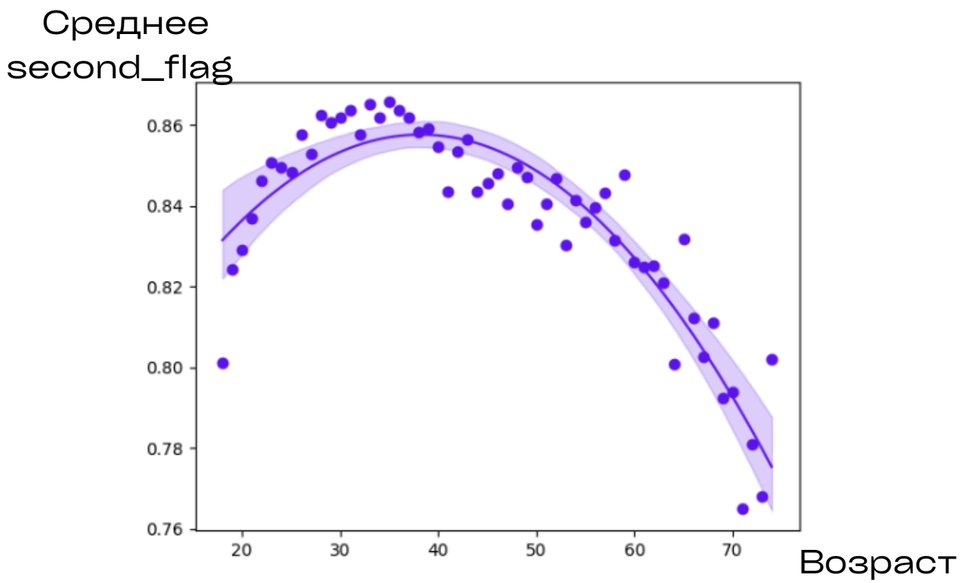


Проверка возможности использования логреги

- таргет является двоичным
- наблюдения независимы
- При исследуемом показателе мультиколлинеарности нет, она есть лишь при age и age^2, что логично:

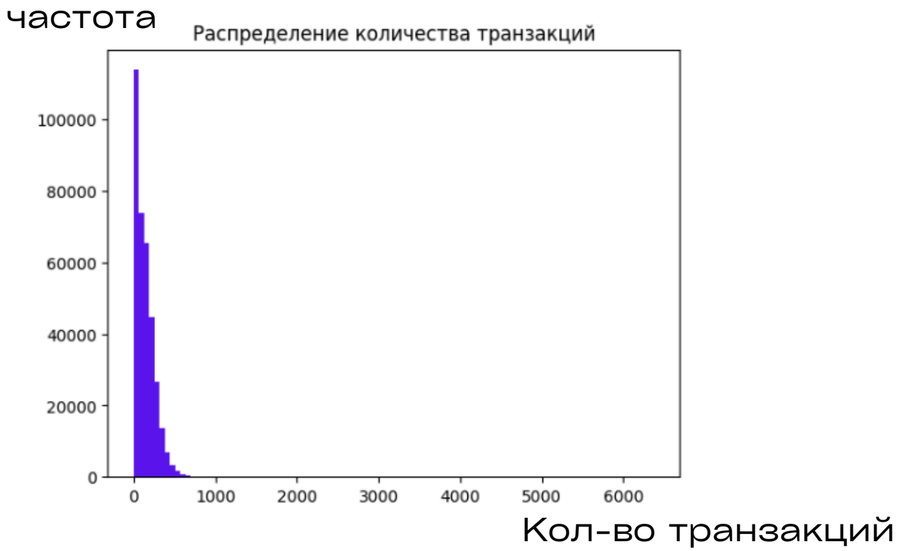
	VIF	variable
0	98.146722	Intercept
1	1.021498	gender_cd
2	28.541528	age
3	28.606483	age2
4	1.041192	transactions_cnt

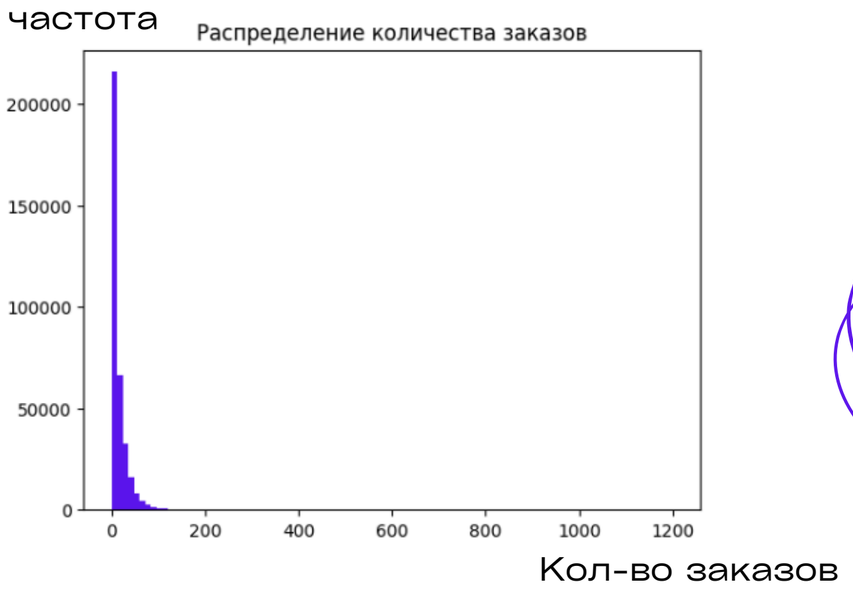
	VIF	variable
0	17.799576	Intercept
1	1.021251	gender_cd
2	1.027262	age
3	1.038883	transactions_cnt

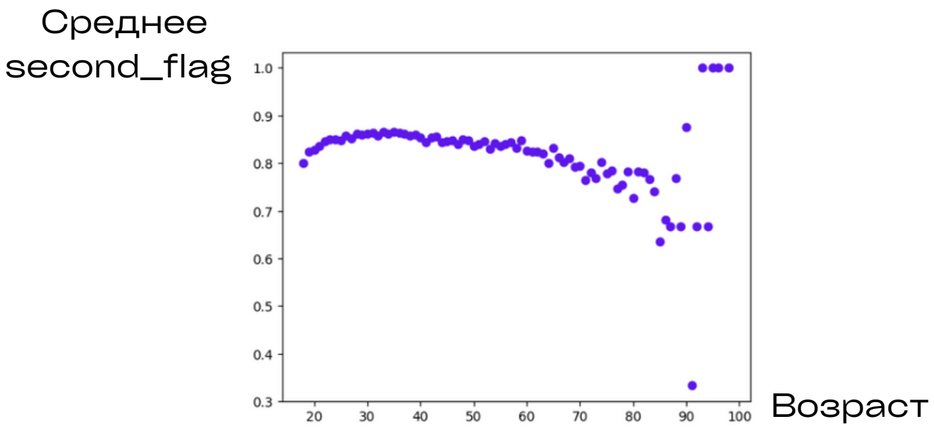


- у переменной age наблюдается четкая квадратичная зависимость с лояльностью, поэтому для того, чтобы это учесть, мы также добавили переменную age^2
- (график обрезали до 75, т.к. наблюдений > 75 крайне мало, поэтому возникают экстремальные средние значения (0 и 1)

Необрезанные графики







Условия по другим картам

До 10% баллами

Дополнительно за покупку топлива на АЗС «ЛУКОЙЛ»





До 10% кэшбэк

Бонусами за покупки на АЗС



Строки - дубликаты

- В нашем датасете были как строки - полные дубликаты, так и строки, которые дублировались во всем, кроме записей в Городе - в одних были заказы, а в других нет

- Мы решили удалить дублированные строки без заказов, т.к. маловероятно, что ошибка в программе продублировала и создала реалистичные записи по Городу большому количеству клиентов, а вот продублировала и удалила записи по городу - вполне реалистичная ситуация



Корреляция Спирмена

