Команда мечты

Структура данных

38500+ сессий

2622 оплаченных заказов

1900+ клиентов

нет ни одного клиента, который совершает свою первую сессию

9 месяцев (01.01-09.30.2025)

26 признаков

	Дата/время	ID	Счётчики	Количественные	Категориальные
Число признаков	2	2	2	3	12

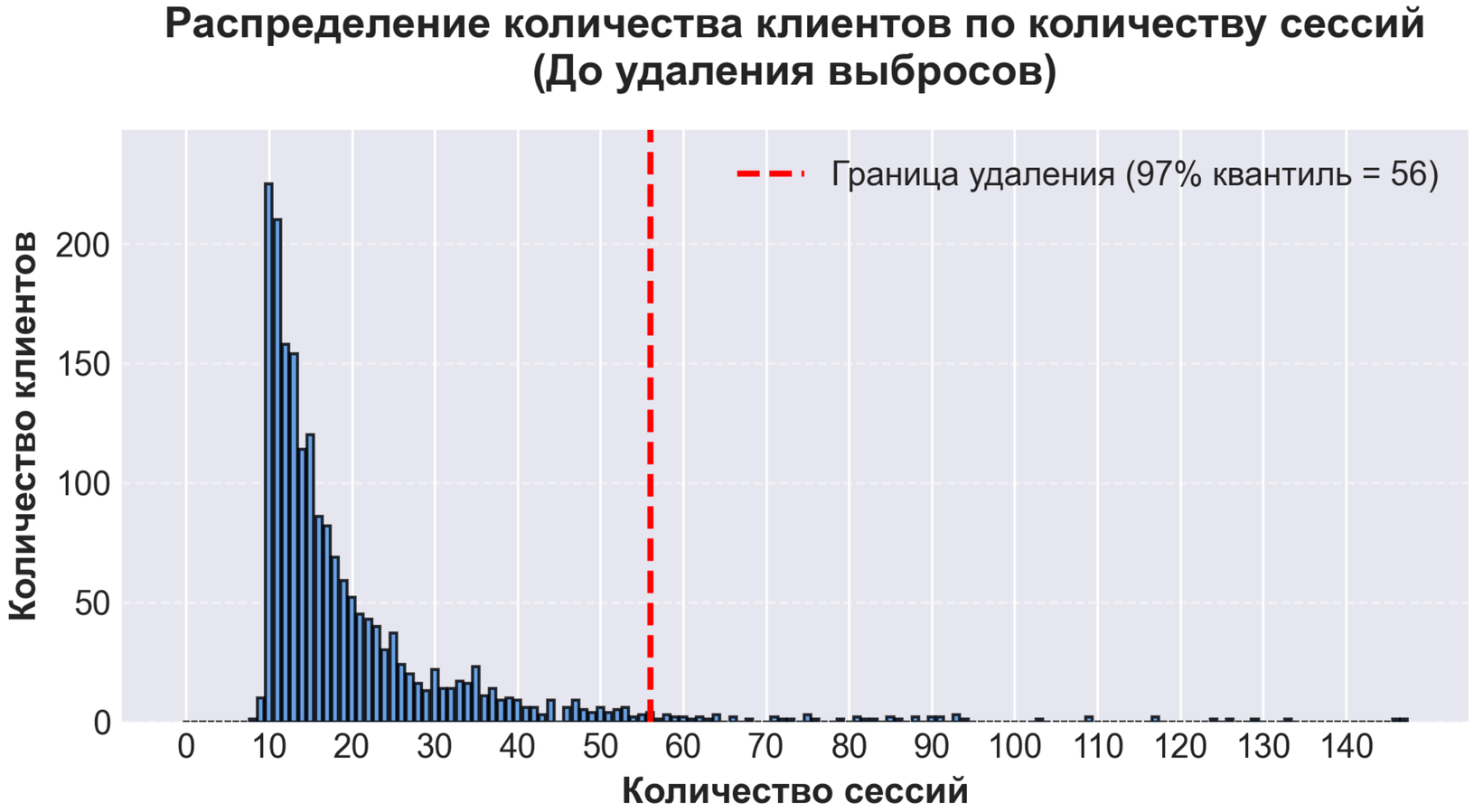
Пропуски в данных:

Знаем информацию о:

- **количестве билетов в заказе**
- **наличии багажа**
- **классе билета**
- **цене билета**
- **точке назначения и отправления**
- **авиалинии**

Только для оплаченных заказов!

Распределения признаков



Важное наблюдение!
Сессии не независимы, паттерны пользователя ⇒ внутрикластерная корреляция по client_id

Теперь поведение малого количества пользователей не будет искажать результаты нашего исследования

Основные конверсии сервиса по дням

Решение:
не использовать часть данных за июль

Петли - возвраты на предыдущие экраны

Классический маршрут	Поиск > Выдача > Оффер > Чекаут > Покупка бэк > Покупка фронт
Сессии с петлями	Поиск > Выдача > Оффер > Выдача > Оффер
	Выдача > Оффер > Выдача > Поиск > Выдача

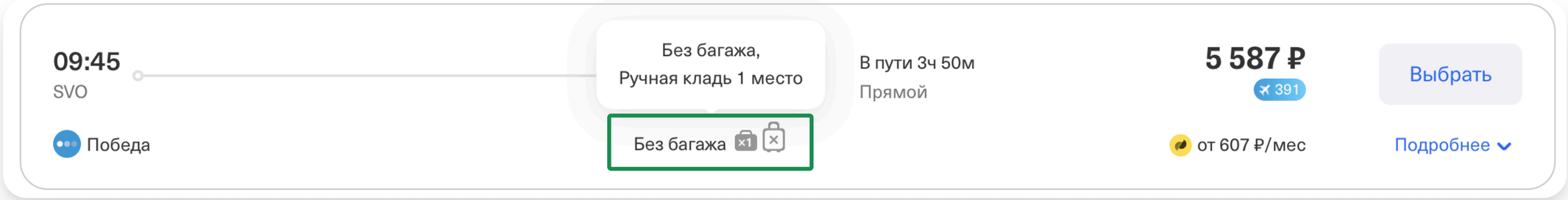
+ информация

-время и оптимальность взаимодействия

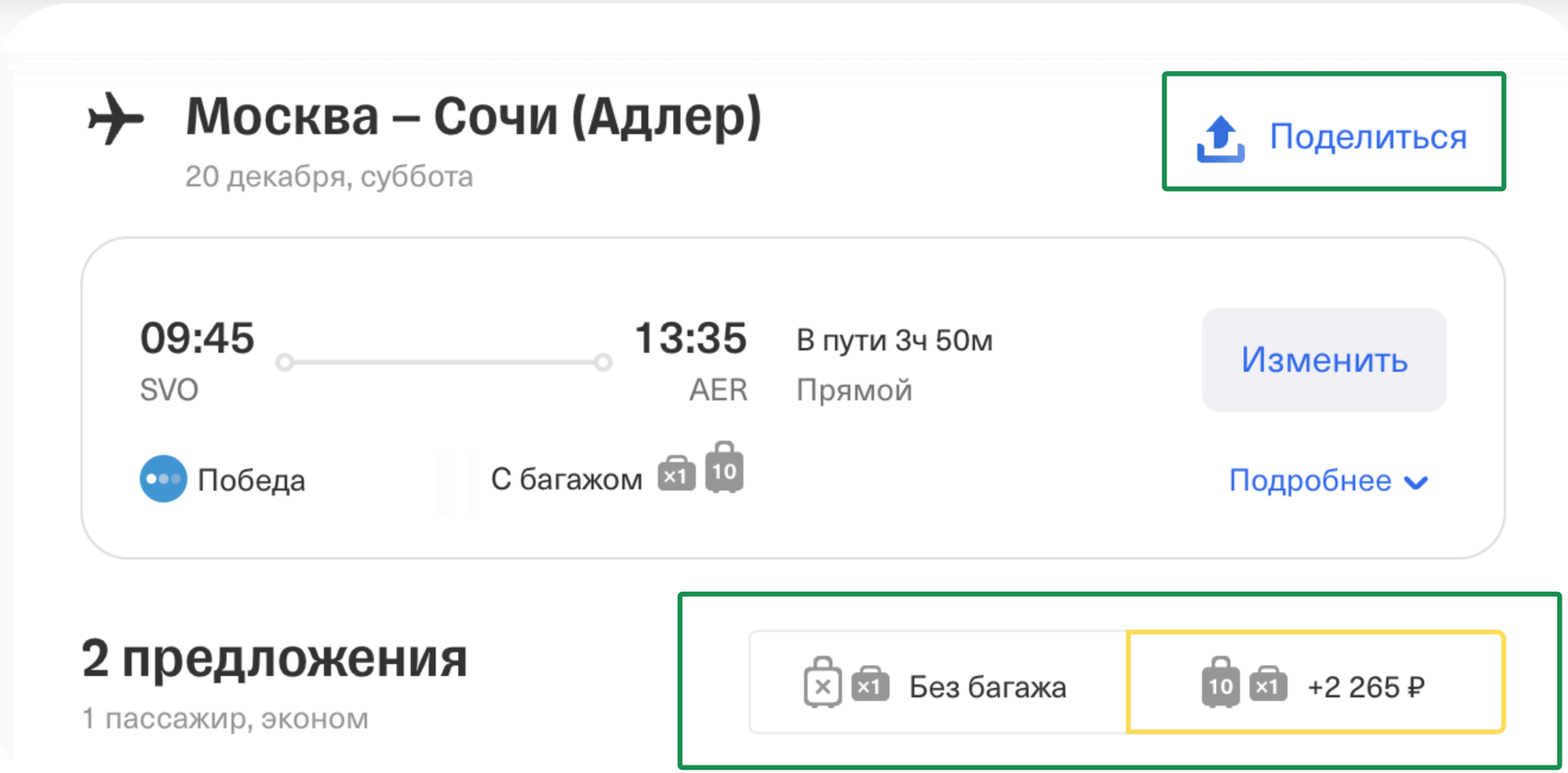
Почему эта зависимость может существовать в данных?

Сайт Т-Авиа

Выдача



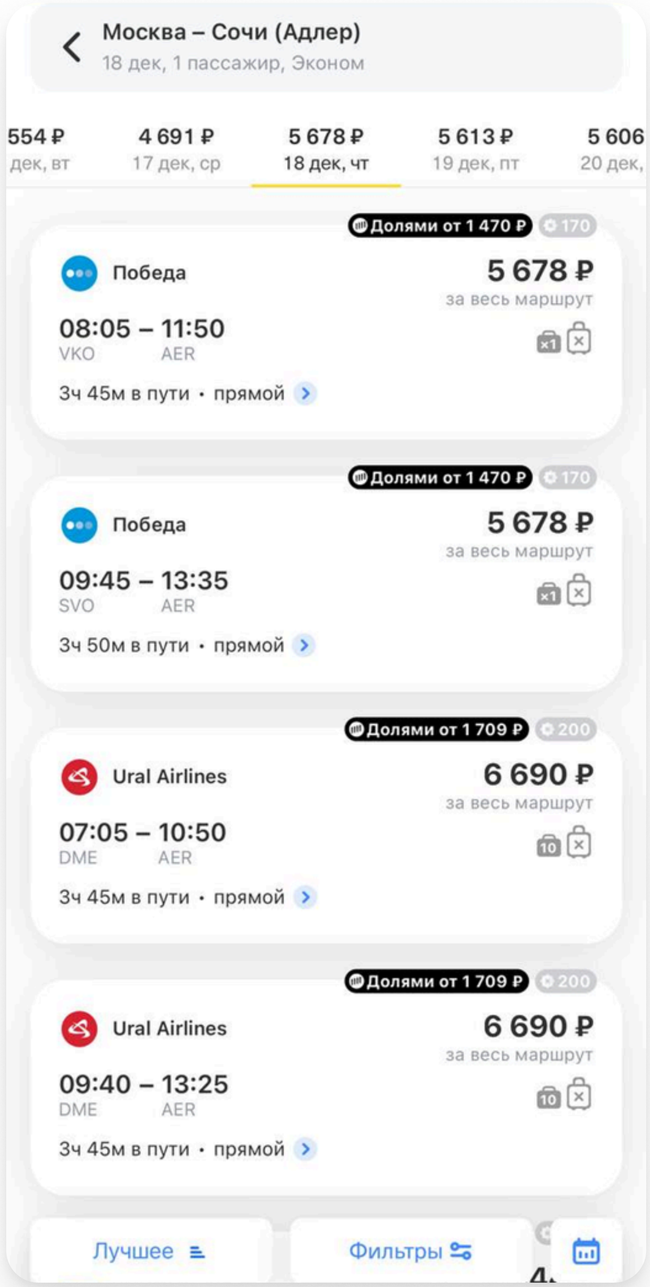
Оффер



при сравнении двух рейсов возникает петля!

Почему эта зависимость может существовать в данных?

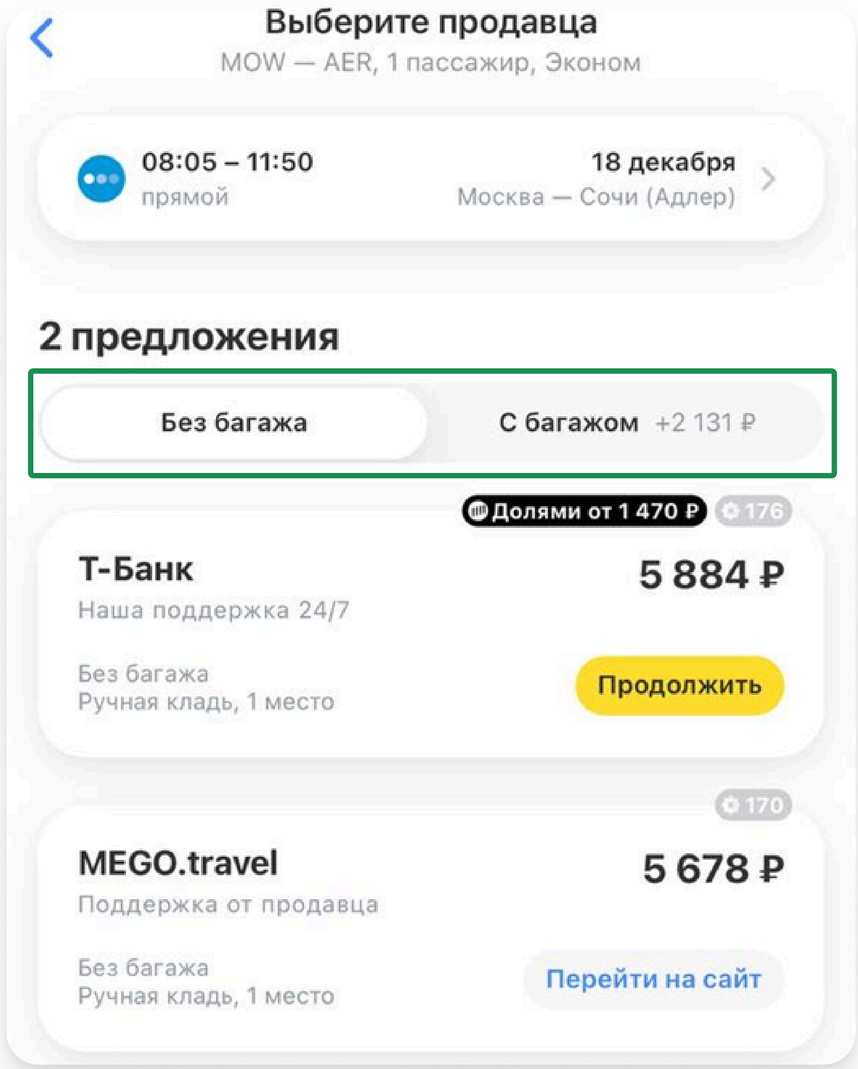
Выдача



Мобильное приложение Т-Авиа

при сравнении двух рейсов возникает петля!

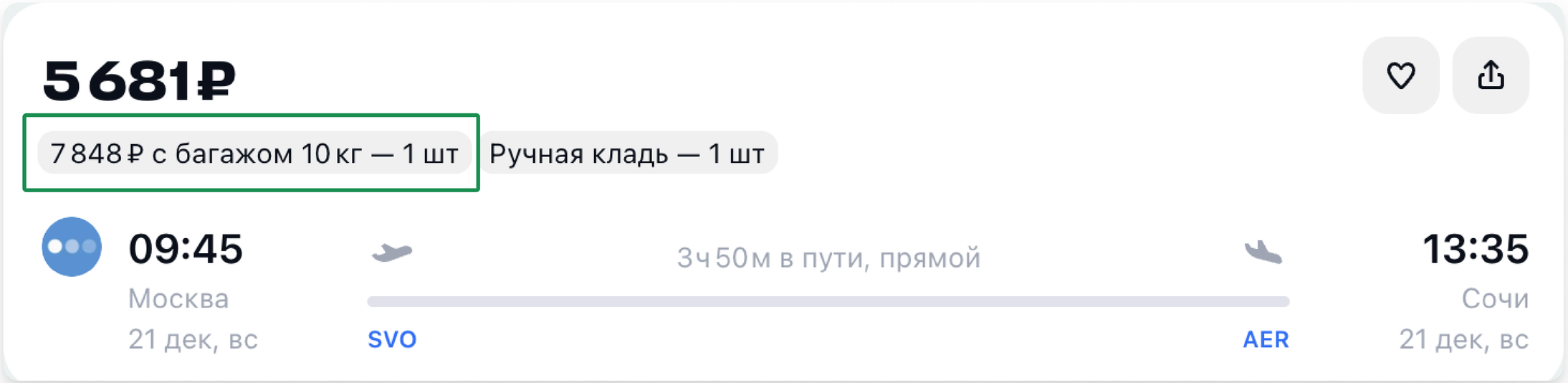
Оффер



Почему эта зависимость может существовать в данных?

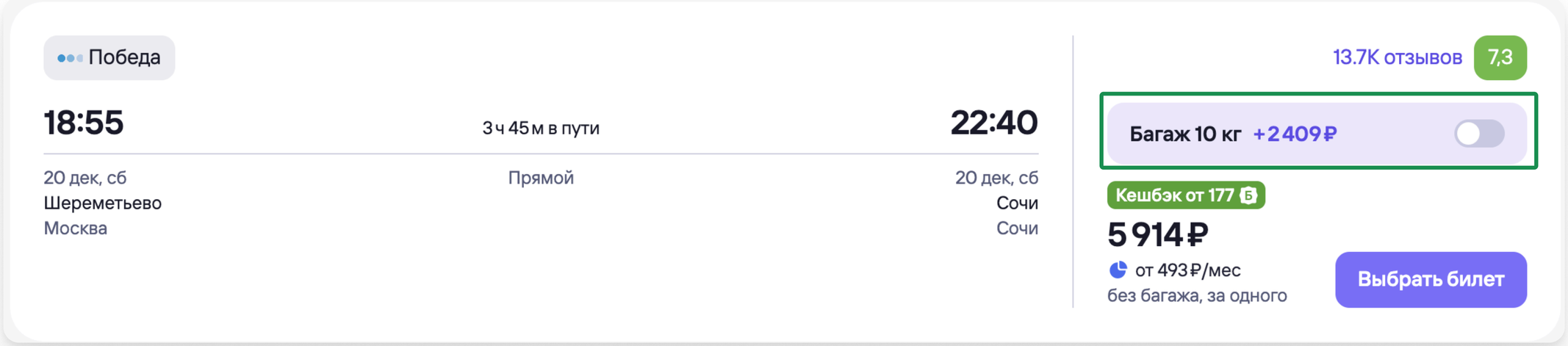
Выдача

Сайт Авиасейлс



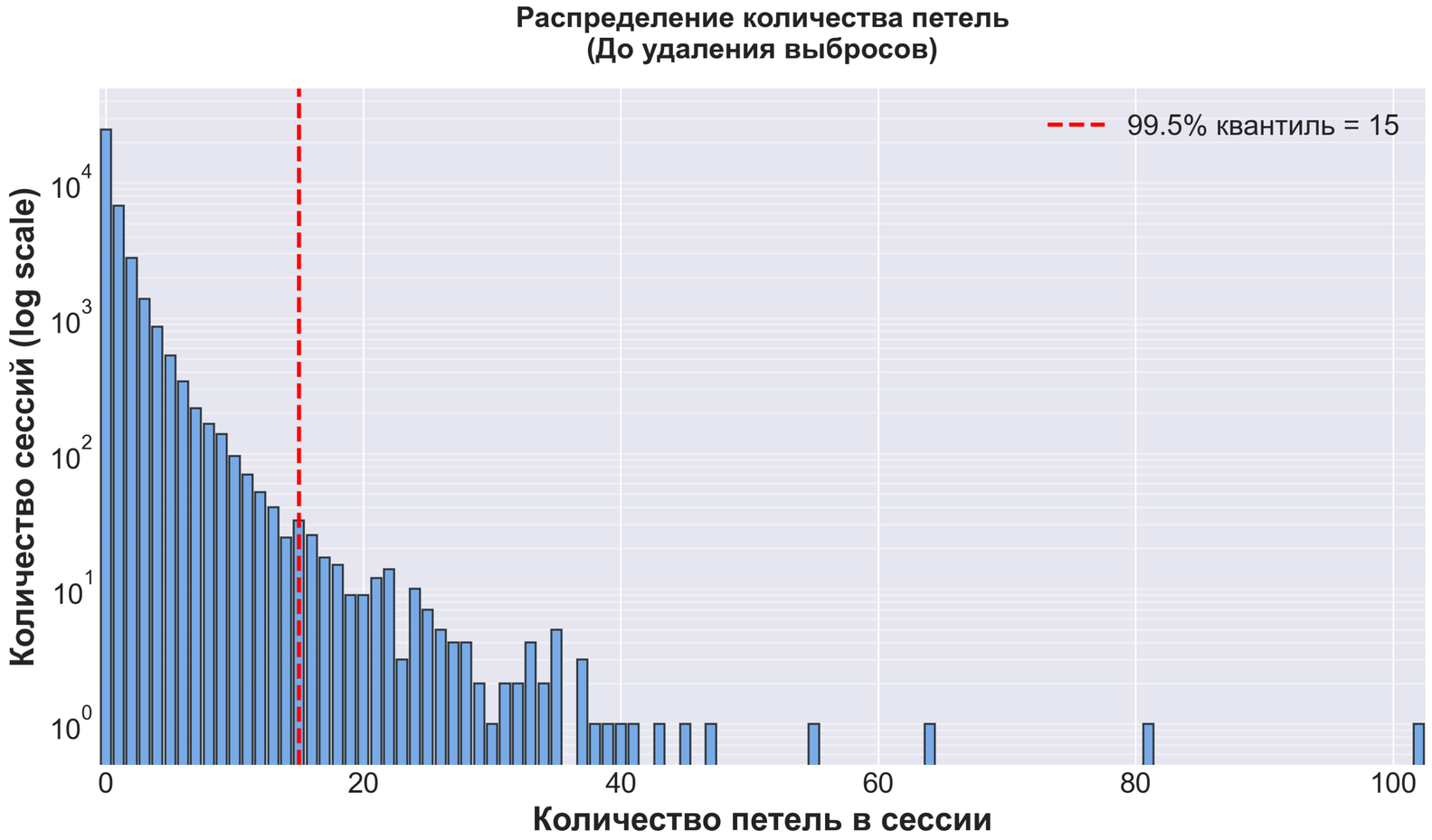
Выдача

Сайт tutu.ru



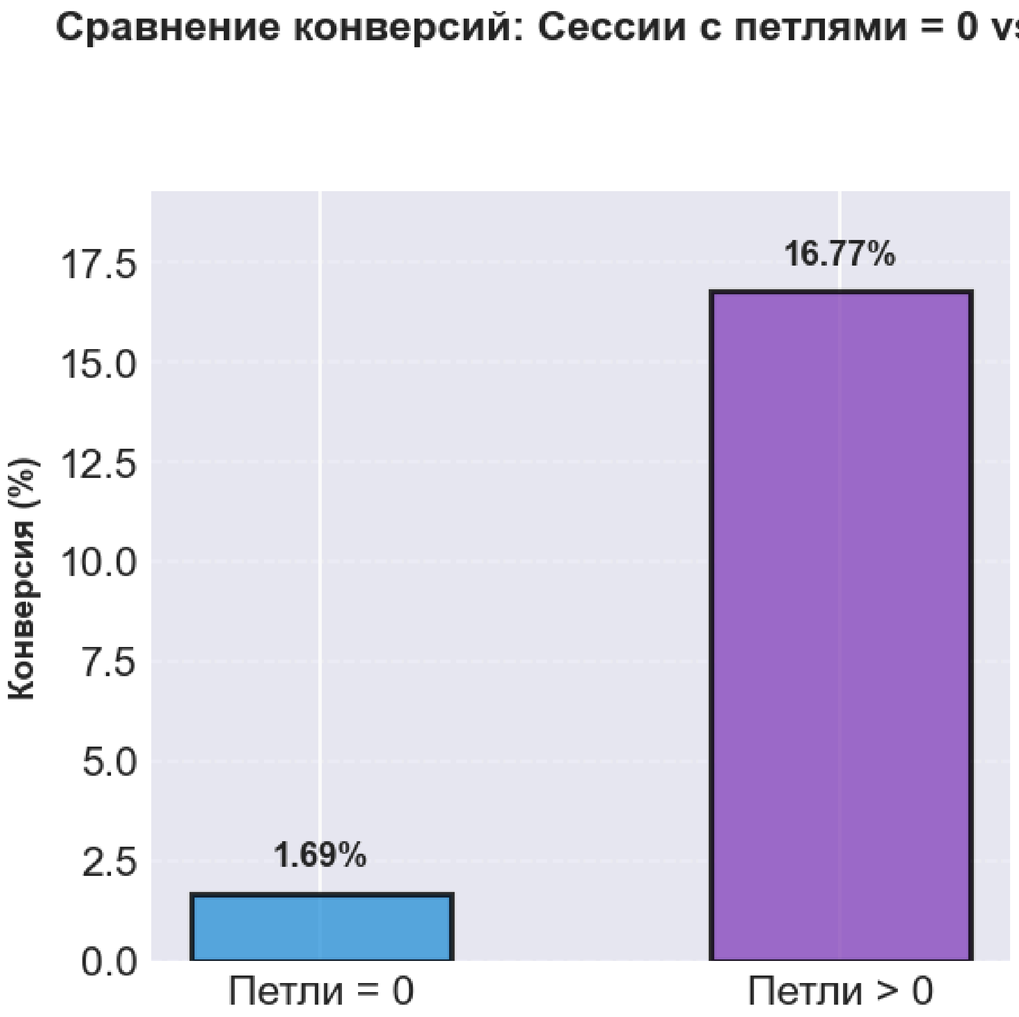
Больше информации на 1 экране - не нужно “петлять”

Петли



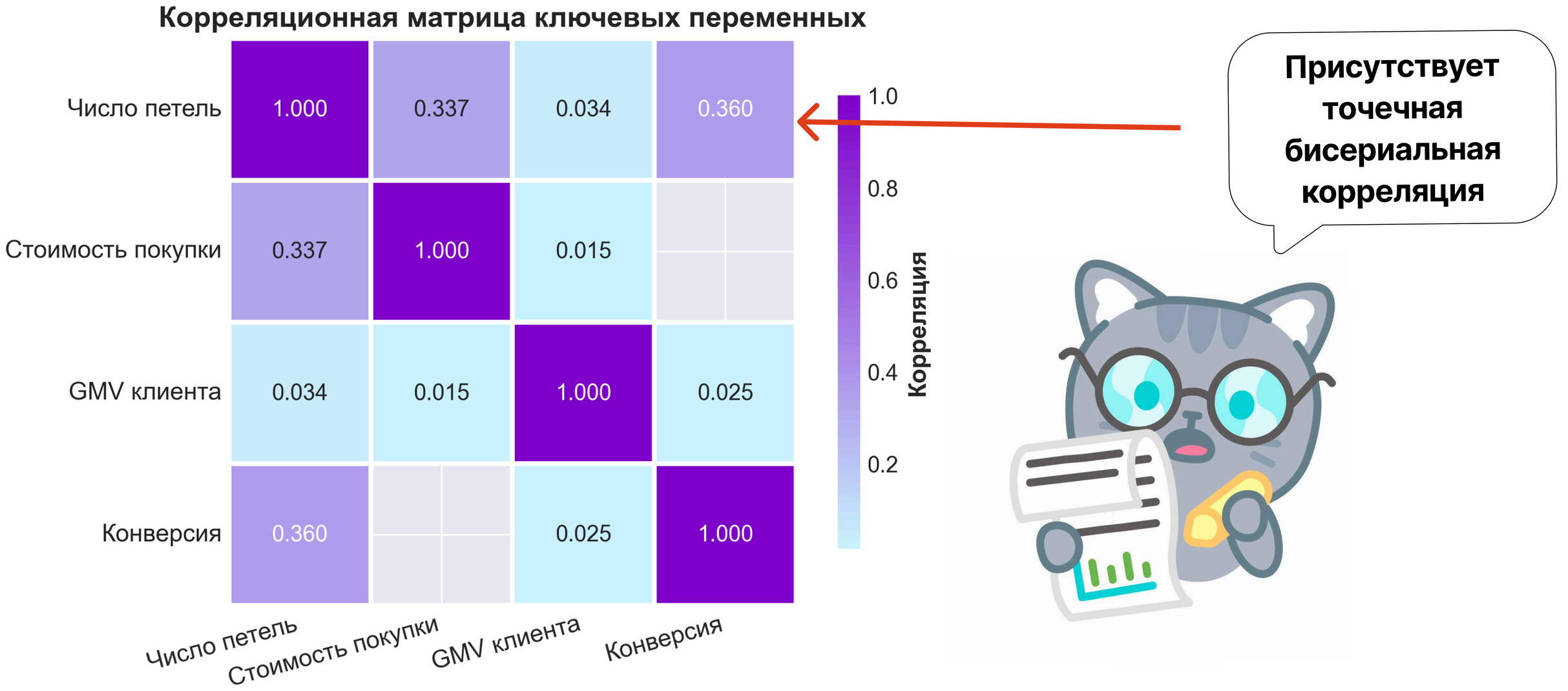
Мы убрали петли выше 0.995 квантиля как выбросы, так как они встречаются слишком редко.

Связь петель и конверсии

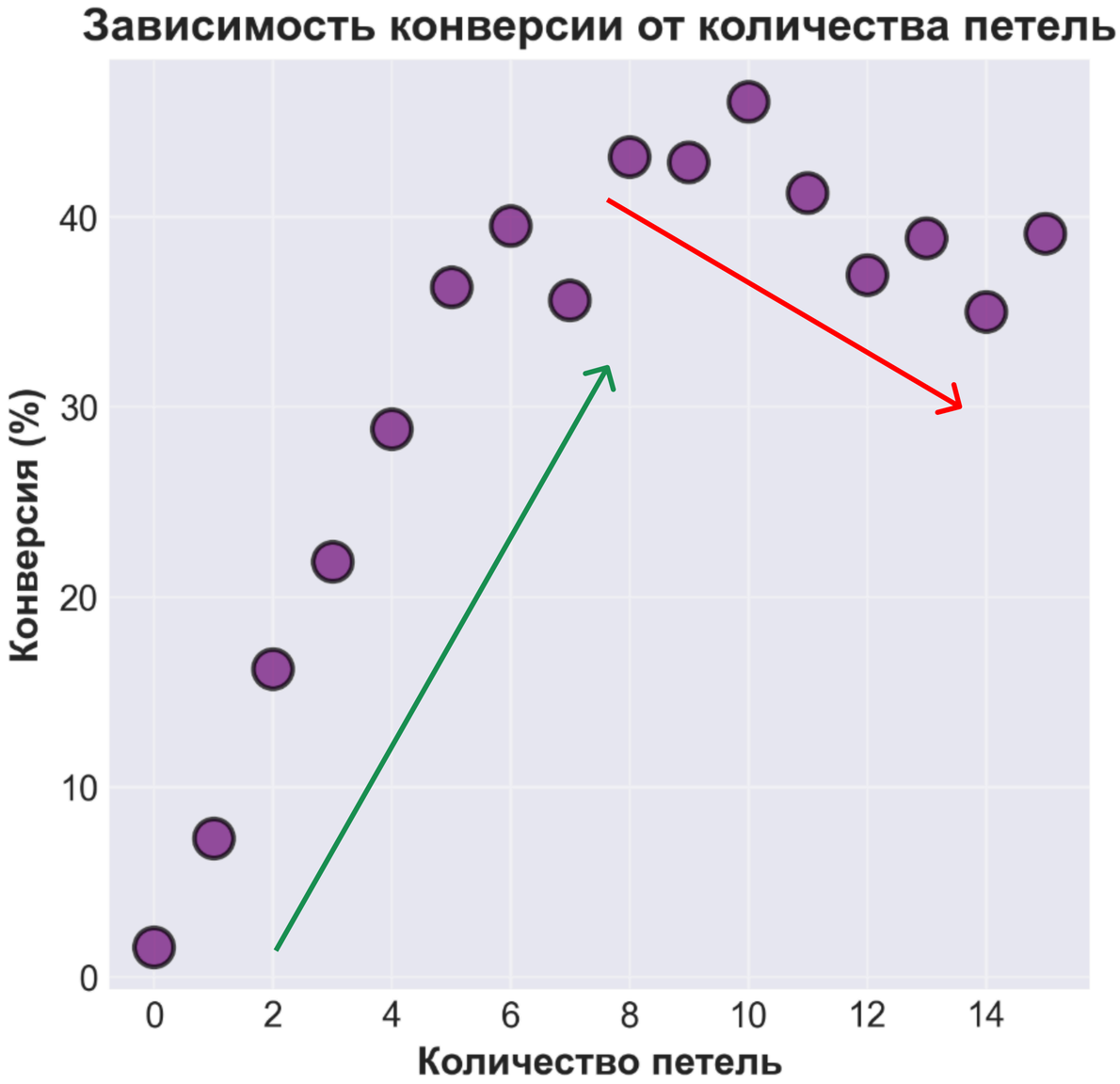


Петли влияют на конверсию в покупку

Анализ корреляций



Присутствует точечная бисериальная корреляция

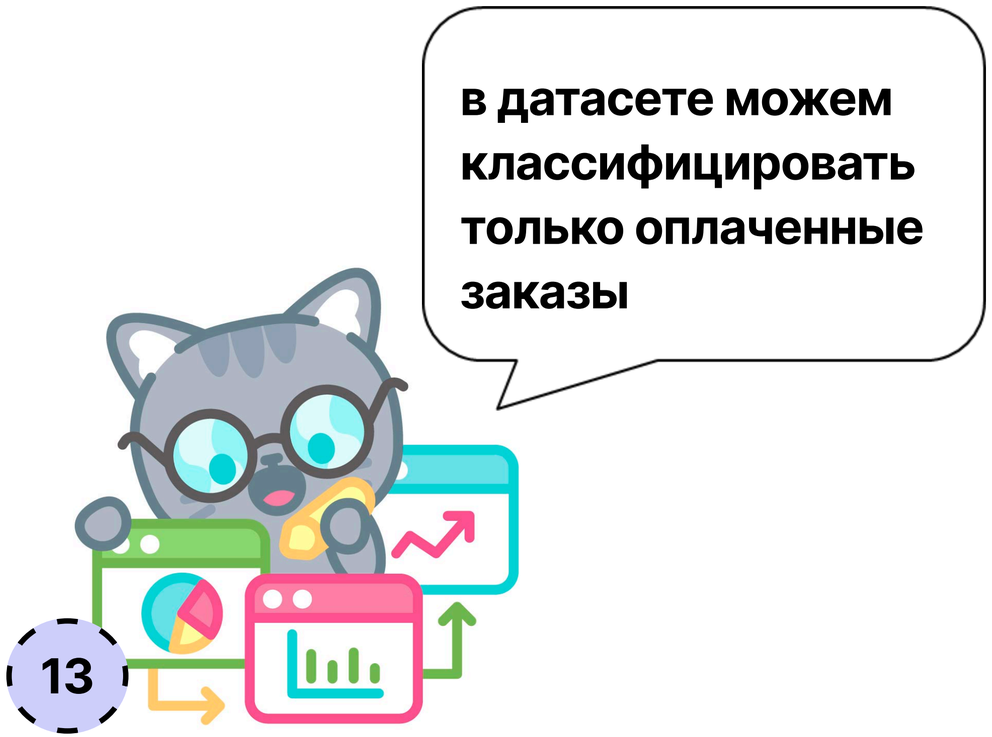


После какого-то момента конверсия начинает падать

Максимальная конверсия в точке - 10 петель

“Простые” бронирования

Пользователя интересует базовый минимум:
-дата и время вылета
-пункт назначения и отправления
-авиакомпания



“Сложные” бронирования

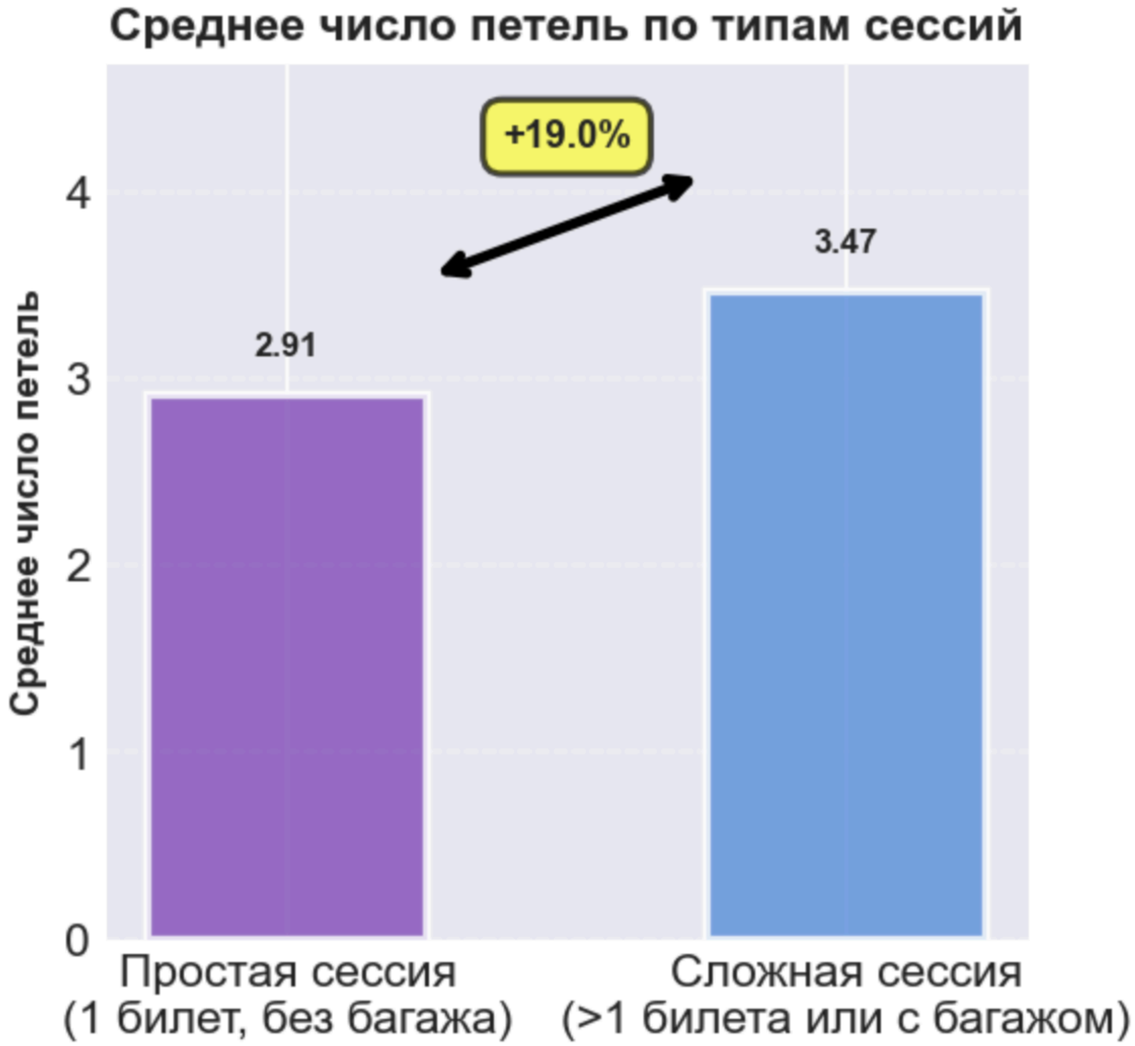
То же самое, но также:
-цена доплаты багажа
-возможность коллективного выбора билетов

```
# В бронировании есть багаж
session_purchase_tickets_luggage_flg == 1

# В бронировании несколько билетов
session_purchase_tickets_num > 1
```

Петли у сложных сессий

Наше наблюдение из опыта присутствует и в данных.
В среднем у сложных бронирований больше петель



Выводы EDA

Описательные статистики	Проанализированы распределения и статистики основных переменных
Выбросы	38561 сессия до очистки → 30310 сессий после очистки
Взаимосвязи	Выявлена нелинейная взаимосвязь между количеством петель и конверсией, а также связь количества петель со “сложностью”
Вывод	Возможно, сложность бронирования связана с конверсией

Исследовательский вопрос

У каких бронирований конверсия меньше?

Гипотеза

Сложные бронирования имеют меньшую конверсию.

МЕХАНИЗМ

Петли связаны с конверсией

↓

Пользователь устаёт от переизбытка петель

↓

Меньше вероятность покупки

↓

У сложных бронирований конверсия меньше

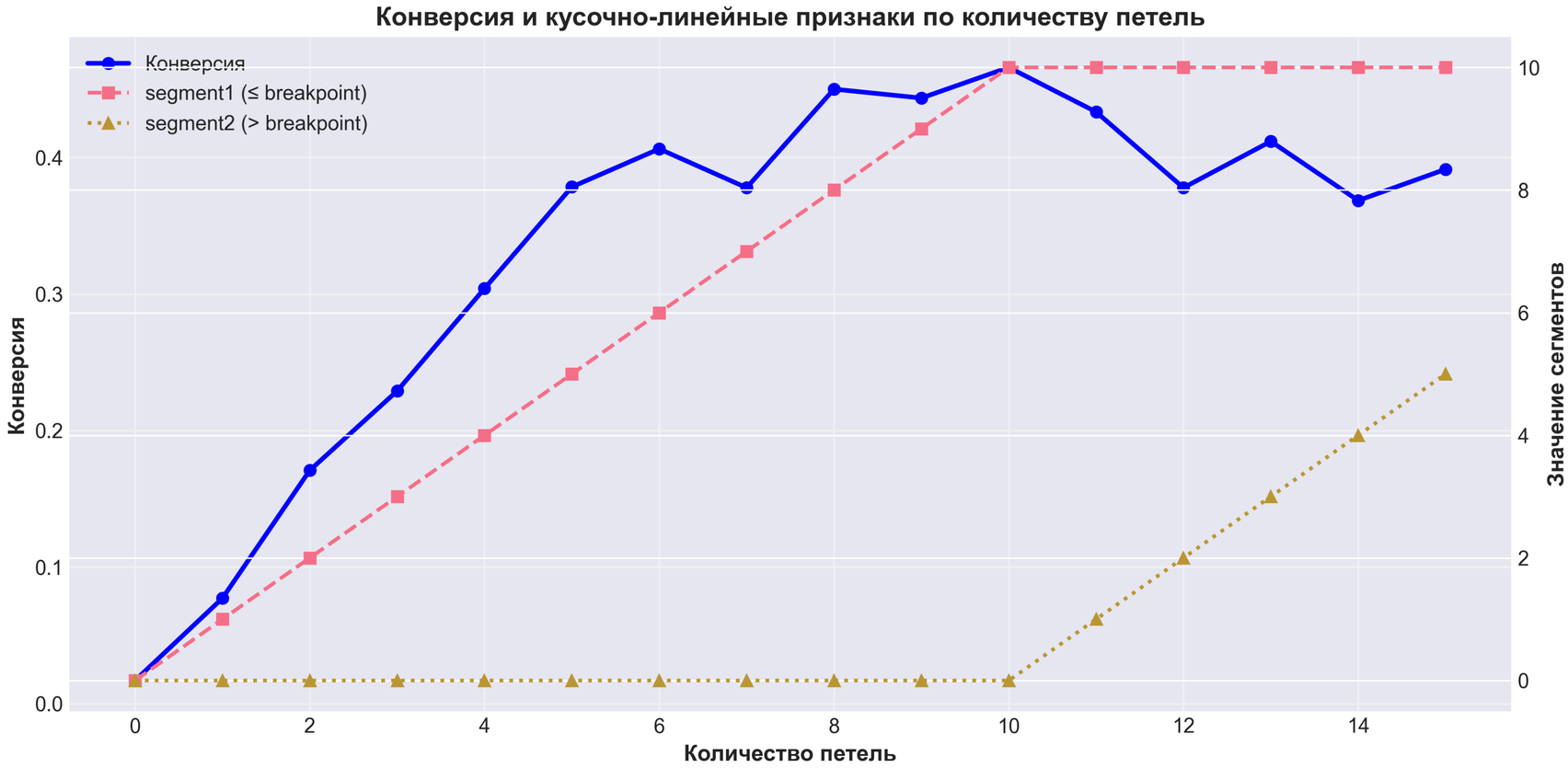
У сложных бронирований больше петель

↓

Линейно-кусочная логистическая регрессия

Мы используем логистическую регрессию с двумя кусочными переменными, созданными для разных областей зависимости

Точка перегиба - 10 петель



$$\text{logit} = \alpha + \beta_1 f_1 + \beta_2 f_2,$$

Линейно-кусочная логистическая регрессия

требования мат. метода

Мультиколлинеарность отсутствует ✓
Но наблюдения зависимы ×
→ модель можно использовать с поправкой на корреляцию внутри кластеров клиентов

Решение:

cluster-robust Логистическая регрессия

Обоснована ли линейно-кусочная модель?

Действительно ли в данных есть точка перегиба?

H0 — **Коэффициенты не различаются, точка перегиба отсутствует**

H1 — **Коэффициенты различаются, точка перегиба присутствует**

Для проверки используется Wald-тест

α = 0.05

Действительно ли это спад?

Коэффициент в зоне спада < 0, и он статистически значим

H0 — **Коэффициент в зоне спада не значим, зависимость отсутствует**

H1 — **Коэффициент в зоне спада значим, присутствует убывающая зависимость**

$\alpha = 0.05$

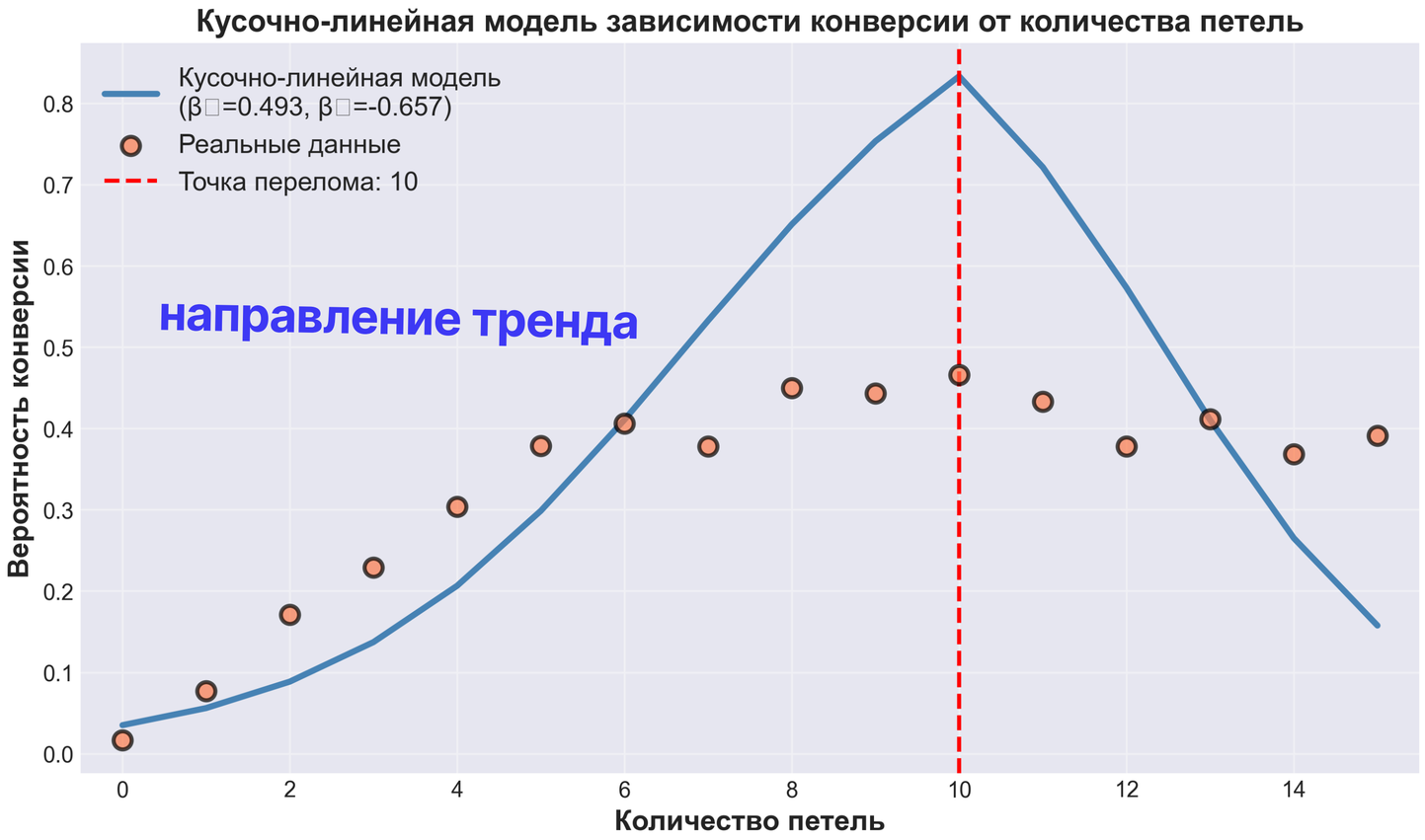
Логистическая регрессия

Для проверки различия коэффициентов мы использовали Wald-тест

Коэффициент спада статистически значимо отличается от коэффициента подъёма.

Pseudo R2: 0.166

P-value: < 0.05



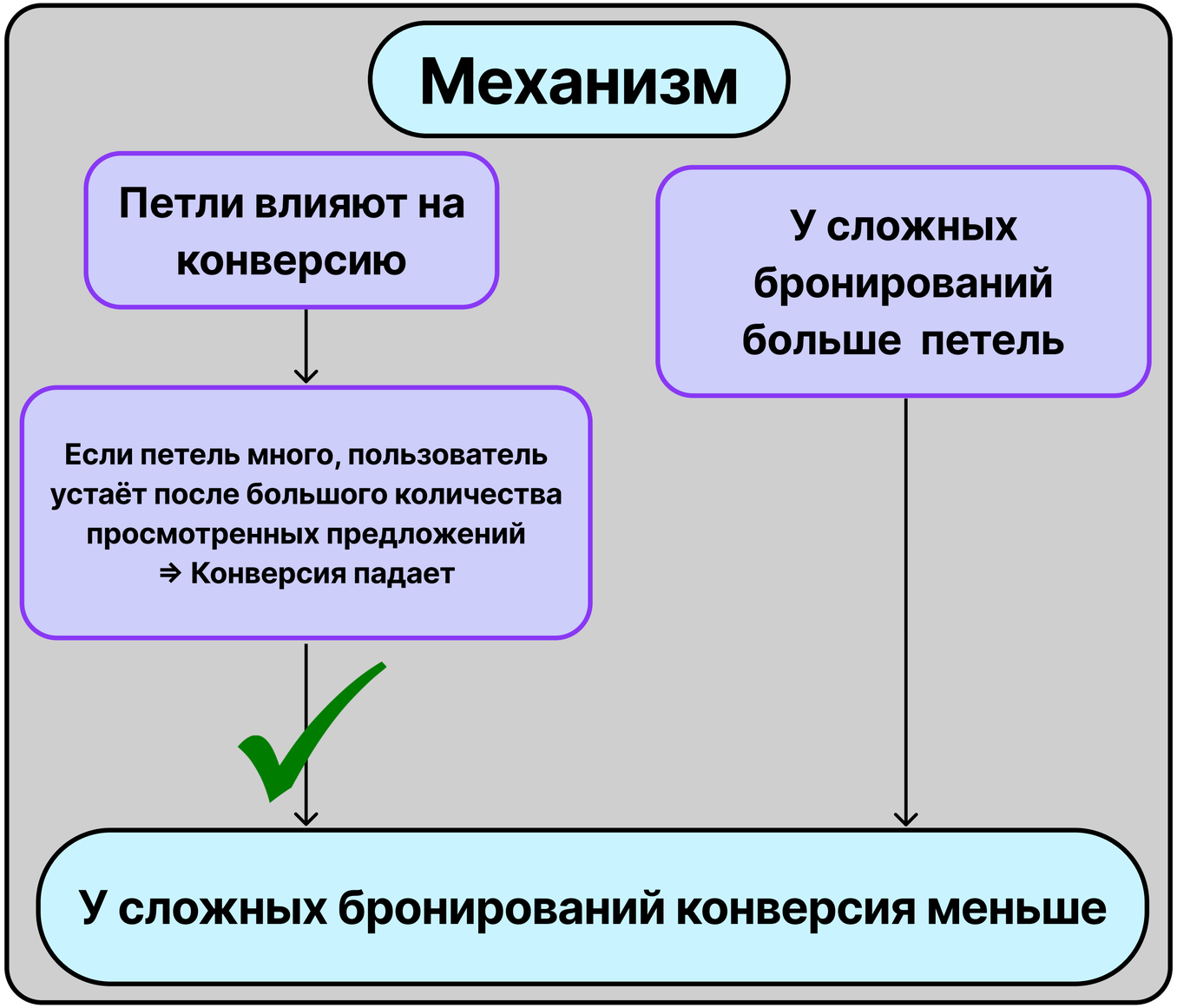
$$\text{logit} = \alpha + \beta_1 f_1 + \beta_2 f_2,$$

Коэффициент в зоне спада < 0 и статистически значим.

P-value: < 0.05

Интерпретация результатов мат-модели №1

Лог. регрессия с кусочно-линейным признаком значимо лучше описывает данные, чем модель без него, она указывает на спад конверсии в покупку после достижения 10 петель.



Правда ли, что у сложных броней больше петель?

! Вспомним про зависимость наблюдений ⇒ нельзя использовать t-test и другие методы

Решение: использовать линейную регрессию с поправкой (cluster-robust)

требования мат. метода соблюдены

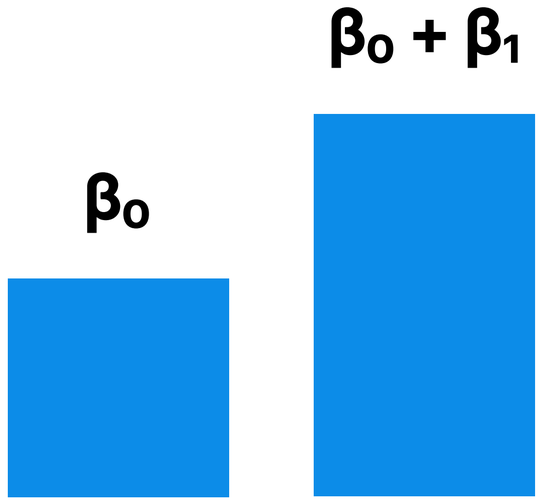
кол-во петель

$$Y_i = \beta_0 + \beta_1 G_i + \varepsilon_i$$

is_hard

при Gi = 0, b1*Gi = 0
⇒ b0 = среднее кол-во петель для простых

при Gi = 1, b1 = среднее кол-во петель для сложных - b0



α = 0.05

⇒ b1 показывает разницу средних, проверяем стат-значимость и величину коэфициента

H0 — Среднее число петель у простых и сложных бронирований одинаково.

H1 — Средние числа петель статистически различаются.

Сравнение средних

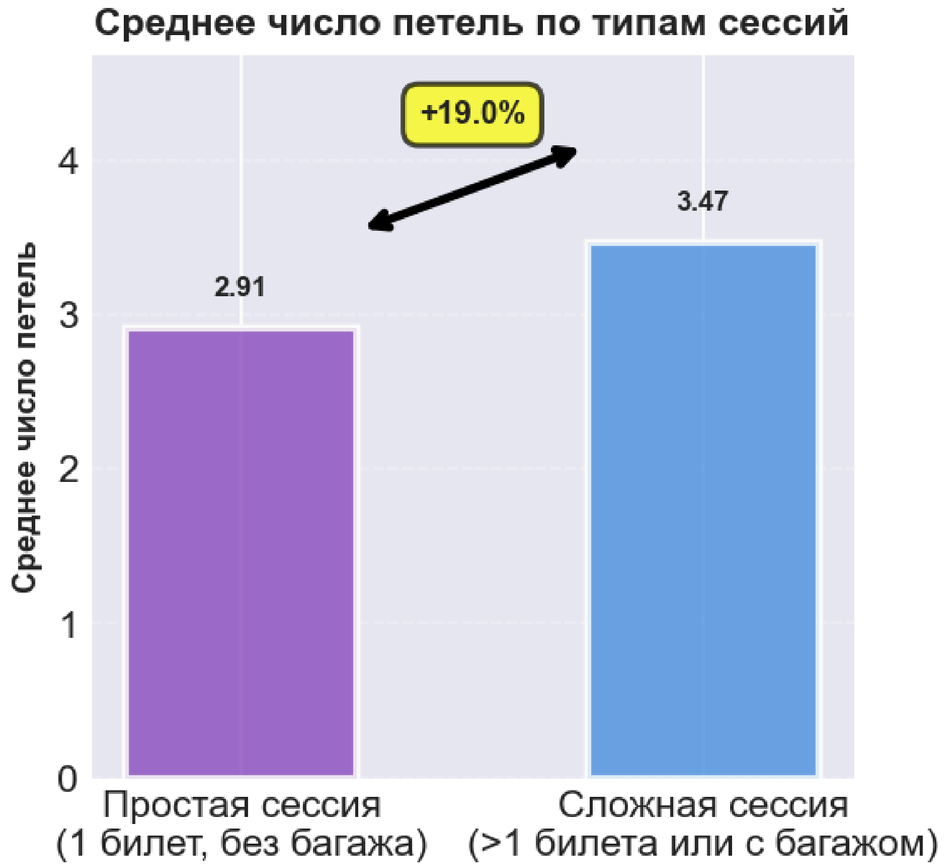
Для проверки различия групп мы использовали линейную регрессию

Простые сессии
1 билет и без багажа

Сложные сессии
>1 билета или с багажом

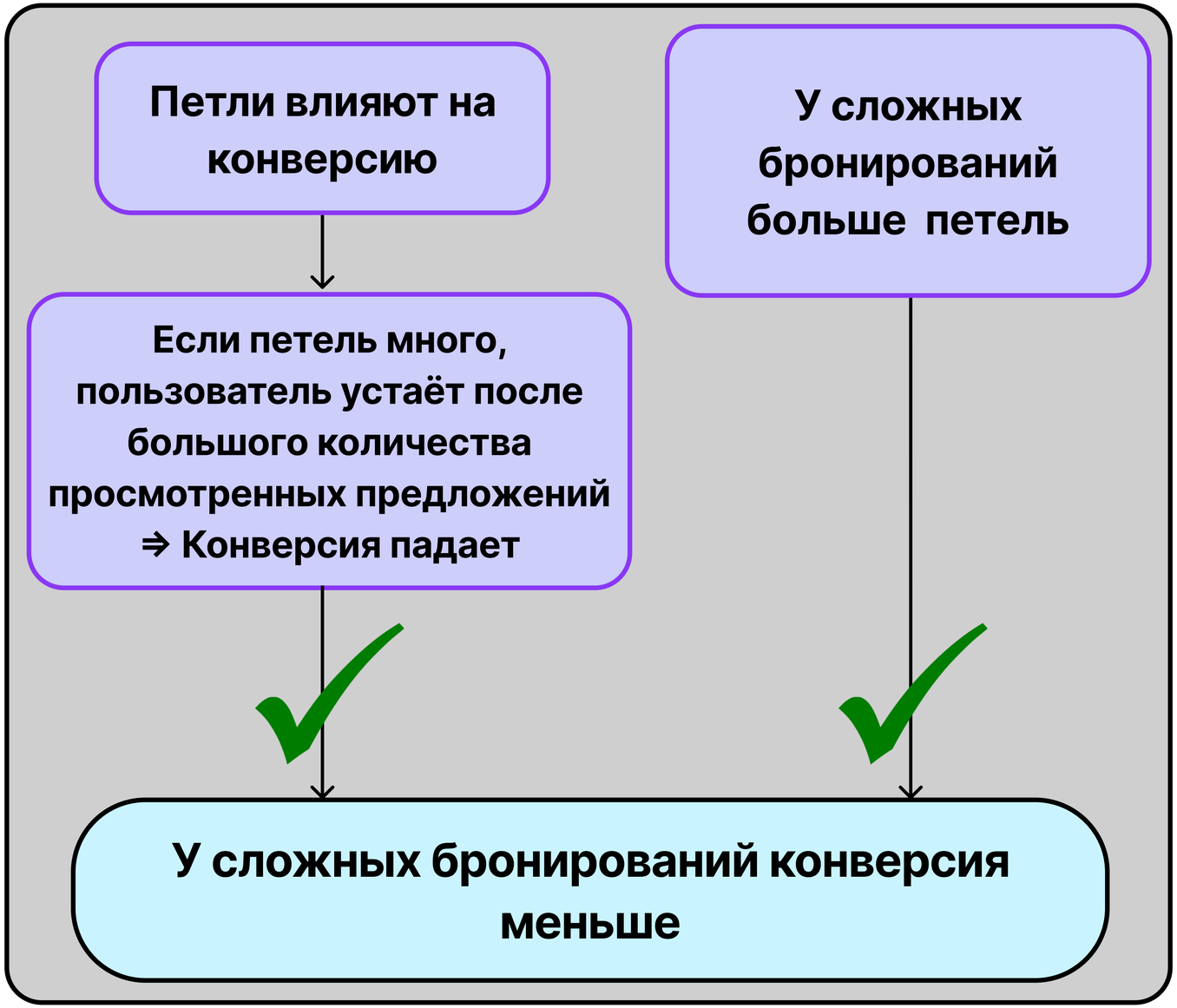
Группы статистически значимо различаются
P-value: < 0.05

Доказательство сравнения средних: Сложные сессии имеют больше петель



Интерпретация результатов мат-модели №2

Группы сложных и простых бронирований статистически значимо различаются по числу петель



Устойчивость модели

$$\text{logit} = \alpha + \beta_1 f_1 + \beta_2 f_2,$$

Таблица результатов моделей по разным сегментам пользователей

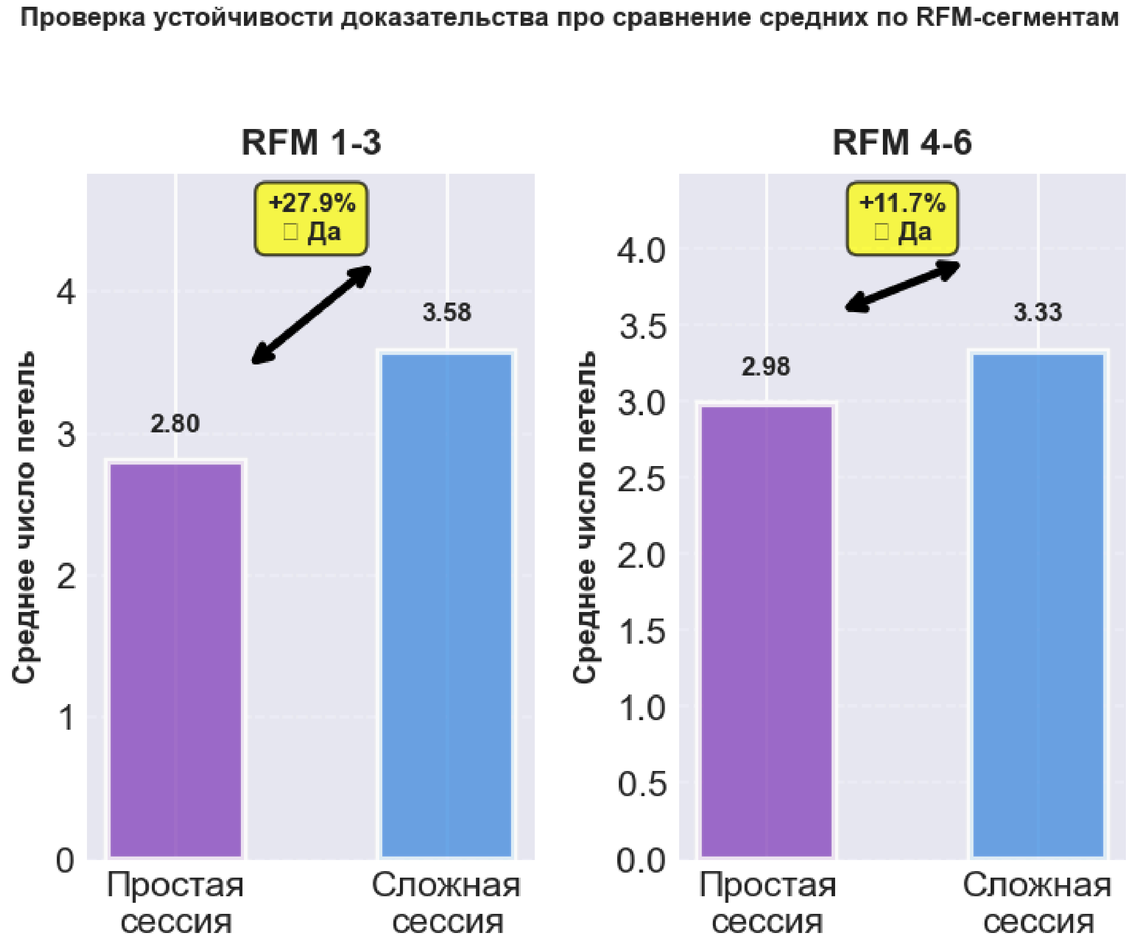
α = 0.05

Подвыборка RFM	β_1 сегмент подъёма	β_1 p-value	β_2 сегмент спада	β_2 p-value	Wald-тест $\beta_2 \neq \beta_1$	Wald-тест p-value
1–3	0.468	0.000	-0.659	0.000	80.648	0.000
4–6	0.521	0.000	-0.665	0.000	93.083	0.000

Мы получаем сравнимо похожие результаты для групп сегментов

Устойчивость модели

α = 0.05



Для проверки устойчивости мат модели мы проверили сравнение среднего количества петель у “сложных” и не “сложных” бронирований в подгруппах rfm <=3 и rfm > 3.

Для группы rfm <= 3:
P-value: < 0.05
Для группы rfm > 3:
P-value: < 0.05

Интерпретация результатов

Спад конверсии от избытка петель
Проверено кусочно-линейной логистической регрессией

У сложных бронирований в среднем выше количество петель
Проверено линейной регрессией

Сложность поведения - количество петель - определяется намерением пользователем, а не фактом финальной покупки. (https://clck.ru/3QoxFB)
валидность перехода подтверждена статьей

Экстраполируем наблюдаемое у успешных бронирований для всех бронирований

У сложных бронирований конверсия меньше

Policy implication

Распознавание «сложного» бронирования до покупки

с помощью ML алгоритма по :
-числу петель
-частым открытиям деталей тарифа
-долгому времени на одном экране, частым фильтрам

Проактивная помощь

Мягко предлагать помощь:
-быстрый чат / ассистент
-быстрое сравнение всех выбранных ранее вариантов
Особенно важно снизить нагрузку на клиента в критический момент

Упростить пользовательский интерфейс

Добавить отображение стоимости багажа в карточке билета (экран выдачи)

Возвращать в воронку ушедших пользователей

-Высылать пуш-уведомления о более выгодных предложения

Альтернативный механизм

Исследование Iyengar & Lepper (2000) показало, что “слишком большой выбор” может демотивировать и снижать вероятность покупки по сравнению с более ограниченным набором вариантов. Академическое основание эффекта choice overload/decision paralysis

Ограничения

Результаты нашего исследования применимы с несколькими ограничениями

Временной промежуток	Новые клиенты	Социально демографические признаки
Нам дан датасет с данными за 9 месяцев	В датасете отсутствуют новые клиенты	Отсутствуют любые соц-дем факторы

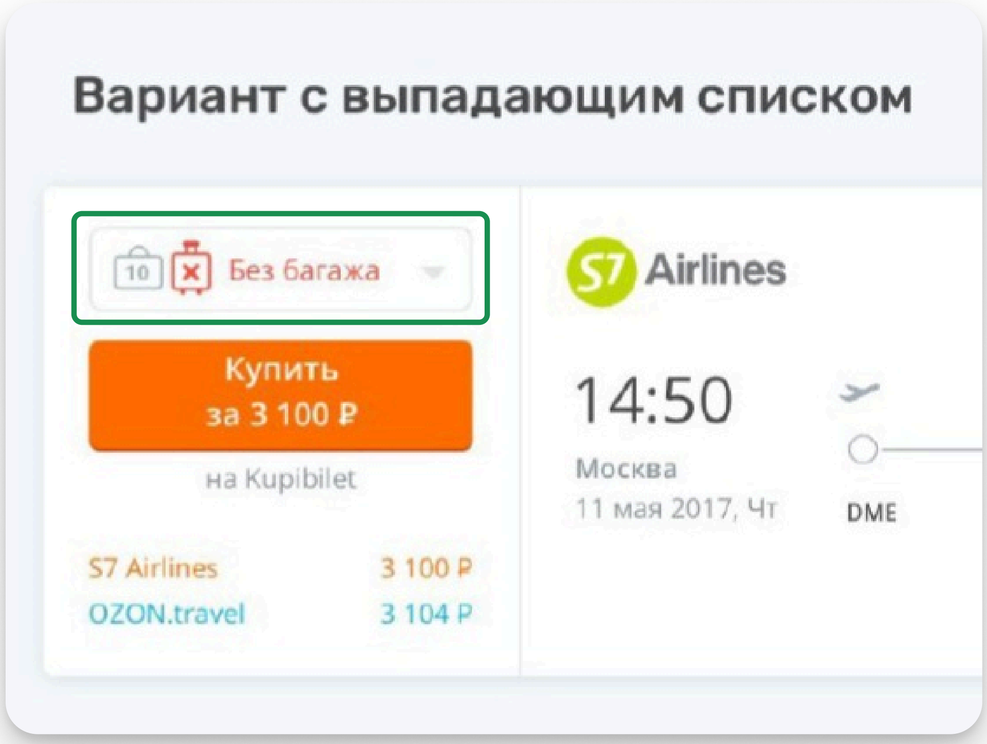
Похожие исследования

- Модель “Search Gaps and Consumer Fatigue” (working paper Ursu, Zhang, Honka) рассматривает длительность и интенсивность поиска по сайтам до покупки и показывает, что рост затрат на поиск (по времени и числу просмотренных опций) повышает “усталость” и приводит к откладыванию покупки.

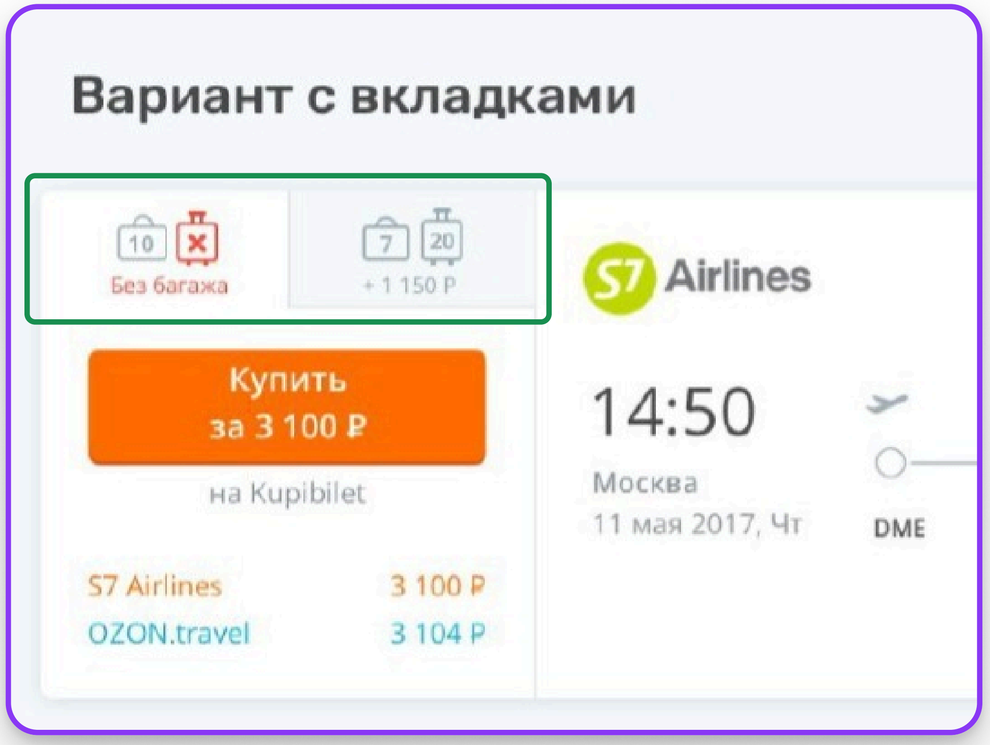
- В данных Baymard Institute по причинам abandonment отдельно фигурирует причина “too long / complicated checkout process” (слишком длинный/сложный процесс покупки), из-за которой заметная доля покупателей бросает корзину.

Похожие исследования

Исследование Aviasales



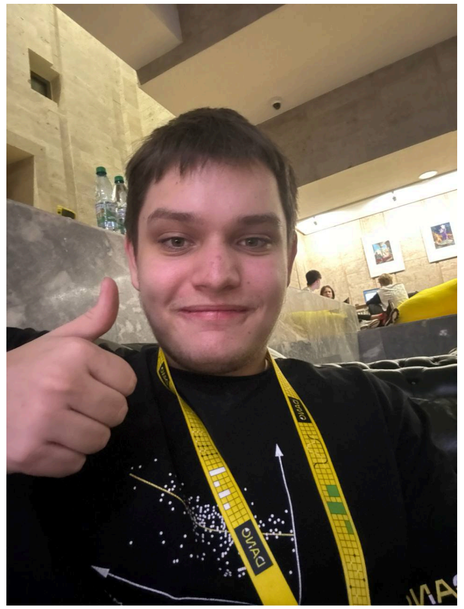
сейчас у Т-Авиа



лучший по результатам того исследования (+как мы предлагаем)

https://clck.ru/3Qpw8o

Наша команда



Около 1% данных находится за точкой перегиба

Около 5% покупок находятся за точкой перегиба

```
================================================================================
ПРОВЕРКА МУЛЬТИКОЛЛИНЕАРНОСТИ (VIF)
================================================================================

Интерпретация VIF:
  • VIF ≤ 5: низкая мультиколлинеарность
  • 5 < VIF ≤ 10: умеренная мультиколлинеарность
  • VIF > 10: высокая мультиколлинеарность (требует внимания)

Результаты:
Переменная   VIF Мультиколлинеарность
  segment1 1.150     Низкая (VIF ≤ 5)
  segment2 1.149     Низкая (VIF ≤ 5)
```

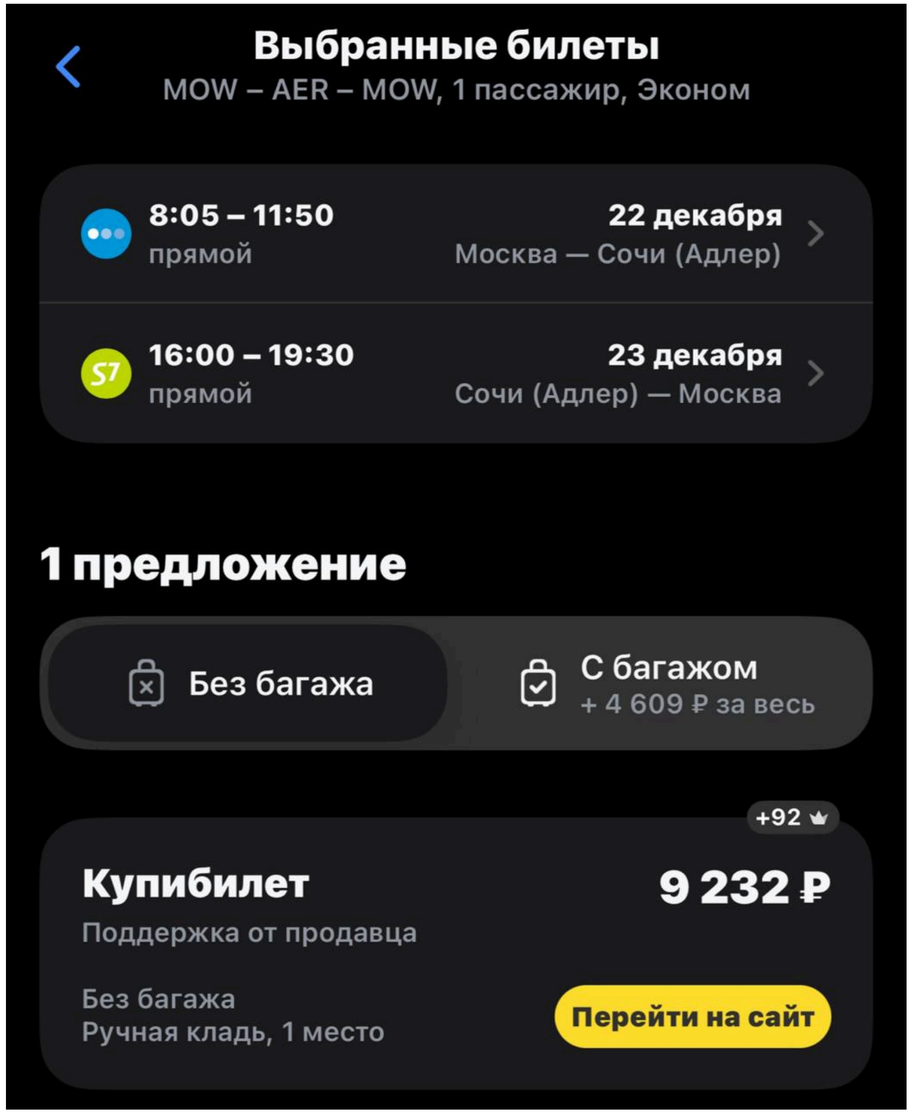
Петли

Плюсы данной метрики:

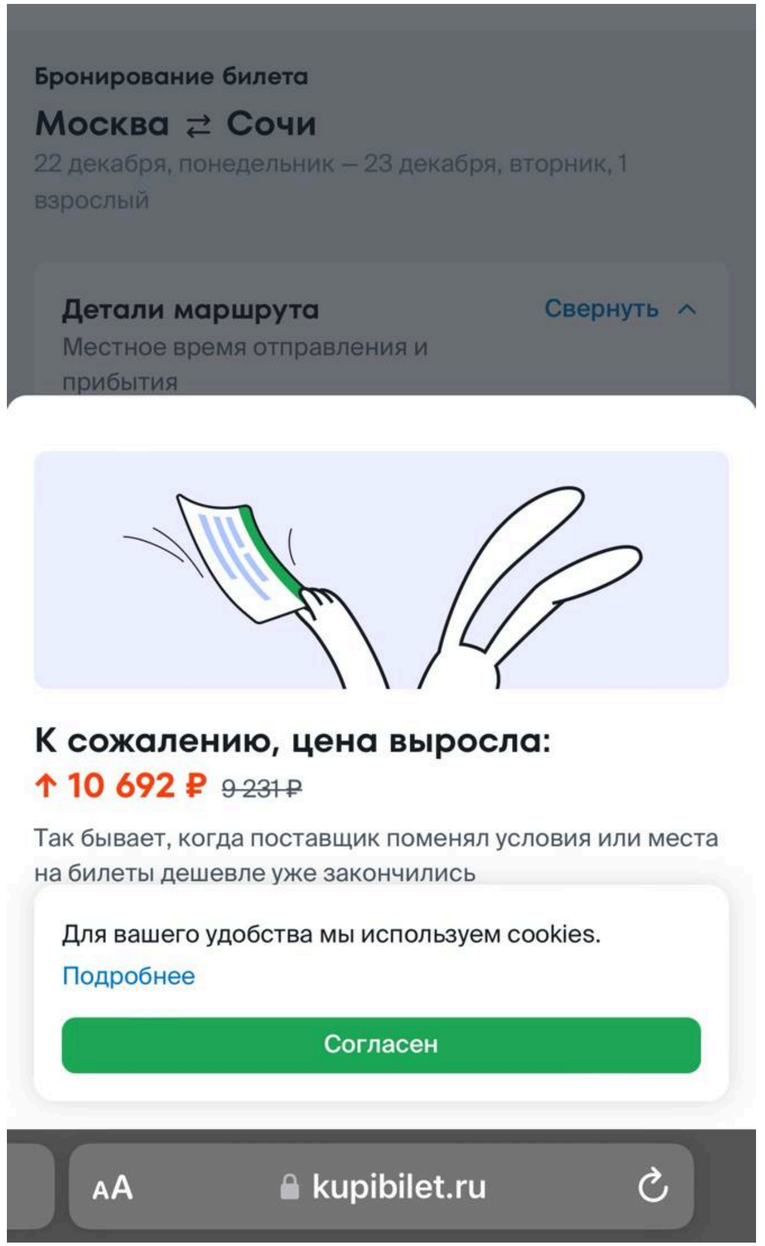
- **отклонение в шагах от кратчайшего пути бронирования**
- **количество избыточных действий на пути к покупке**
- **ситуации, где пользователь переходит по ссылкам не влияют на интерпретацию**
- **вычислительно недорогая**
- **несложно реализуема**

Другие примеры петель

Оффер



Чекаут



Тест на стационарность временного ряда

```
================================================================================
ТЕСТ НА СТАЦИОНАРНОСТЬ: ДИКИ-ФУЛЛЕР ДЛЯ КОНВЕРСИИ В ПОКУПКУ
================================================================================

ДО удаления июля:
--------------------------------------------------------------------------------
ADF статистика: -5.141908
P-value: 0.000012
Критические значения:
  1%: -3.455461
  5%: -2.872593
  10%: -2.572660

✓ Результат: Ряд СТАЦИОНАРЕН (p=0.000012 <= 0.05)
  → Отклоняем H0: ряд не имеет единичного корня

ПОСЛЕ удаления периода 10-29 июля:
--------------------------------------------------------------------------------
ADF статистика: -13.473785
P-value: 0.000000
Критические значения:
  1%: -3.457326
  5%: -2.873410
  10%: -2.573096

✓ Результат: Ряд СТАЦИОНАРЕН (p=0.000000 <= 0.05)
  → Отклоняем H0: ряд не имеет единичного корня

================================================================================
```

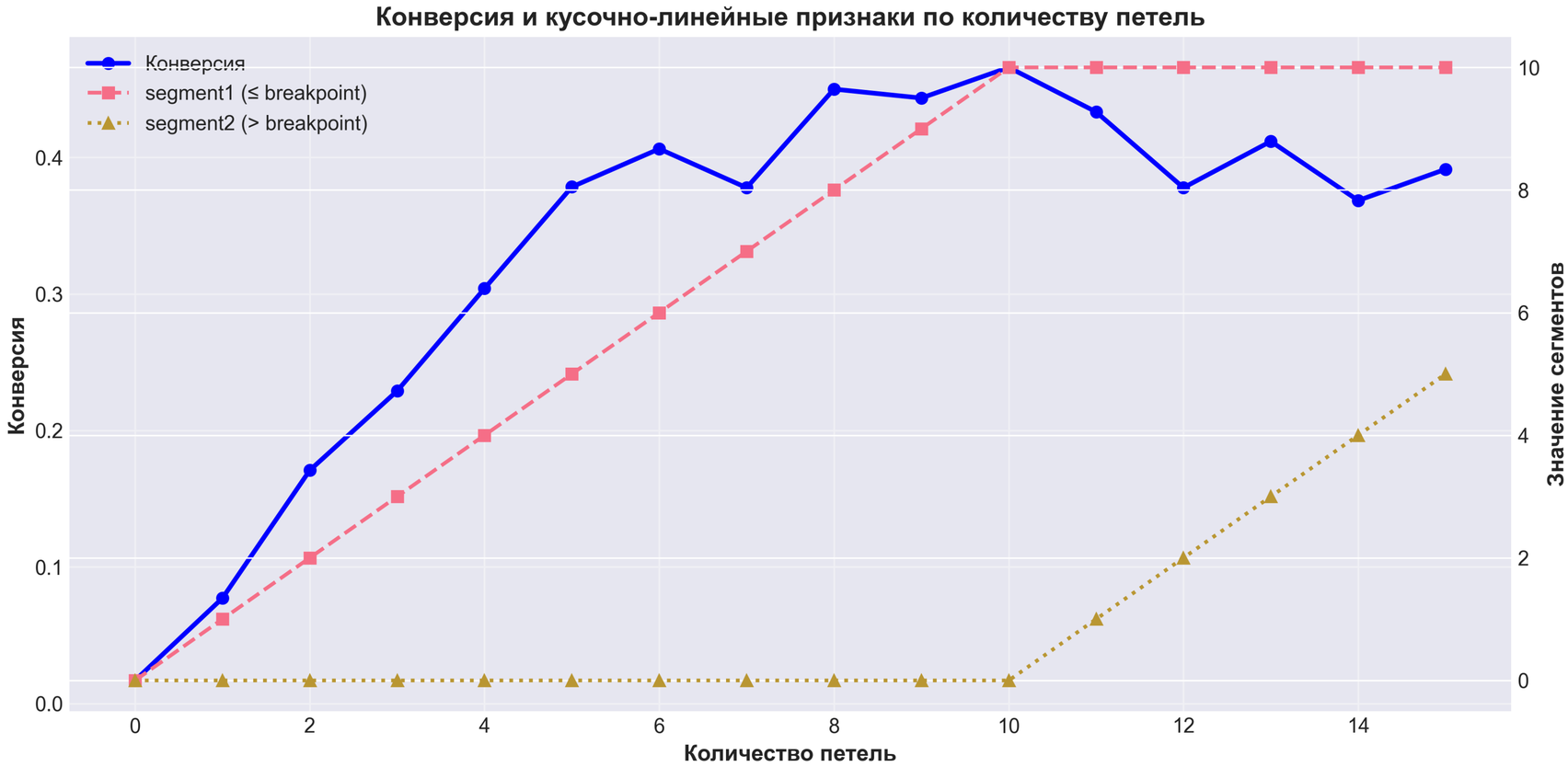
Тест Дики-Фуллера

Пользователь устаёт от переизбытка петель

clck.ru/3Qqnno

Линейно-кусочная логистическая регрессия

$$\text{logit} = \alpha + \beta_1 f_1 + \beta_2 f_2,$$



$$f_1 = \begin{cases} \texttt{num\_loops}, & \texttt{num\_loops} < 10 \\ 10, & \texttt{num\_loops} \geq 10 \end{cases}$$

$$f_2 = \begin{cases} 0, & \texttt{num\_loops} < 10 \\ \texttt{num\_loops} - 10, & \texttt{num\_loops} \geq 10 \end{cases}$$

Приближение спада

Про точечную бисериальную корреляцию

Мы использовали Point-Biserial correlation, потому что хотели оценить связь бинарного признака ‘статус покупки’ с численными метриками. Это корректный способ измерения связи между бинарной и количественной переменной, частный случай корреляции Пирсона.

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}},$$

Про робастность логистической регрессии

Использование cluster-robust стандартных ошибок не изменяет сами оценки коэффициентов модели. Значения коэффициентов (β) остаются идентичными тем, которые получены в стандартной логистической регрессии, поскольку они определяются функцией правдоподобия. Корректировка затрагивает исключительно элементы статистического вывода, а именно:

- Стандартные ошибки (SE) — пересчитываются с учётом корреляции наблюдений внутри кластеров
- z-статистики — изменяются вследствие изменения стандартных ошибок
- p-значения — пересчитываются на основе обновлённых z-статистик
- Доверительные интервалы — расширяются или сужаются в соответствии с робастными SE.

Pseudo-R2

В логистической регрессии используется псевдо-R^2, поскольку классический R^2 неприменим. Псевдо-R^2 отражает улучшение качества подгонки модели по сравнению с нулевой моделью и служит относительной, а не абсолютной мерой объясняющей способности.

Псевдо-R^2 — это мера того, насколько логистическая модель лучше нулевой модели

OLS CLUSTER-ROBUST

Можно, потому что повторяющиеся пользователи не “запрещают” линейную регрессию — они ломают только наивные p-value и доверительные интервалы.

- OLS коэффициенты можно оценивать на данных “сессии внутри пользователей”, если выполняется ключевое условие: в среднем ошибка не связана с регрессорами $E(k|X)=0$
- Проблема в том, что сессии одного пользователя коррелированы, и обычные стандартные ошибки считают их независимыми ⇒ занижают SE и делают p-value слишком маленькими.
- Поэтому мы можем использовать OLS с кластер-робастными (sandwich/Huber–White) стандартными ошибками по user_id, которые допускают любую корреляцию и гетероскедастичность внутри пользователя и требуют независимости только между пользователями.

OLS CLUSTER-ROBUST остатки

коэффициенты считаются приблизительно нормально распределёнными при большом числе кластеров (пользователей), а стандартные ошибки считаются по “sandwich” формуле, допускающей произвольную форму распределения и корреляции внутри user. Что остаётся важным вместо “нормальности остатков”:
независимость между пользователями (кластерами), достаточно много пользователей (иначе p-value/CI могут быть неточными;

“После того как мы разделили сессии по X, в каждой группе не должно быть систематически ‘скрытой причины’, которая ещё дополнительно двигает петли вверх или вниз.

Экзогенность

Экзогенность в регрессии — это требование, что ваш признак X не связан с “неучтёнными причинами”, которые тоже влияют на Y.

$E(k|X) = 0$

То есть: если зафиксировать X, то средняя ошибка равна нулю.

Экзогенность означает:

в группах complexity=0 и complexity=1 эти “прочие факторы” в среднем одинаковы, и не добавляют систематического смещения.

Экзогенность

```
--------------------------------------------------------------------
ТЕСТ 1: КОРРЕЛЯЦИЯ МЕЖДУ РЕГРЕССОРАМИ И ОСТАТКАМИ
--------------------------------------------------------------------

Корреляция complex_session с остатками:
  Коэффициент корреляции: 0.000000
  P-value: 1.000000
  ✓  Статистически значимая малая корреляция (< 0.05) - экзогенность не отвергается
```

Лин. регрессия

```
--------------------------------------------------------------------
ТЕСТ 1: КОРРЕЛЯЦИЯ МЕЖДУ РЕГРЕССОРАМИ И ОСТАТКАМИ
--------------------------------------------------------------------

Корреляция segment1 с остатками:
  Коэффициент корреляции: 0.030137
  P-value: 0.000000
  ✓  Статистически значимая малая корреляция (< 0.05) - экзогенность не отвергается

Корреляция segment2 с остатками:
  Коэффициент корреляции: 0.004284
  P-value: 0.445825
  ✓ Нет значимой корреляции (p >= 0.05) - экзогенность не отвергается
```

Лог. регрессия

Wald-test

Лог. регрессия

$$W = \frac{(\hat{\beta}_1 - \hat{\beta}_2)^2}{\mathrm{Var}(\hat{\beta}_1 - \hat{\beta}_2)}$$

где

$$\mathrm{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \mathrm{SE}(\hat{\beta}_1)^2 + \mathrm{SE}(\hat{\beta}_2)^2 - 2\,\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$