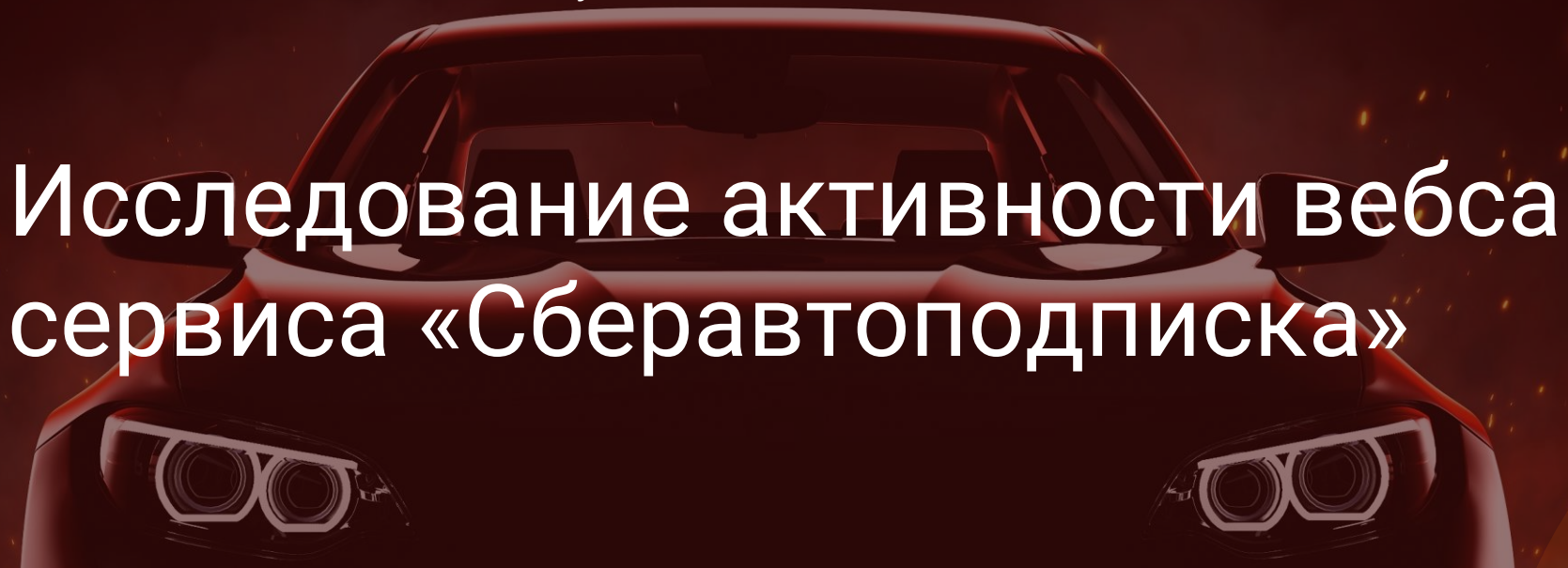
ИТОГОВЫЙ ПРОЕКТ

Курс: “Профессия Data Scientist PRO”
Часть 1 «Введение в Data Science»
Специализация: “Data Analyst”



Исследование активности вебсайта сервиса «Сберавтоподписка»

Луцевич Анна

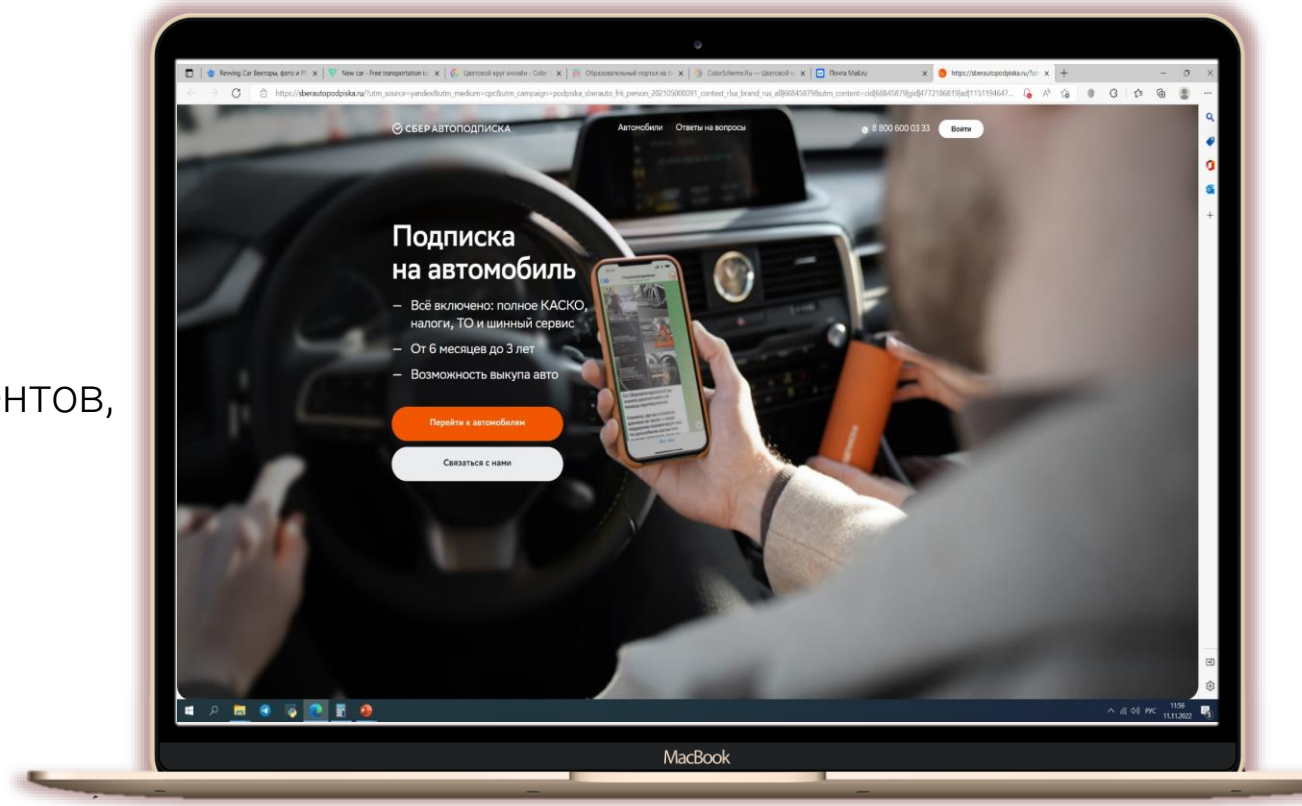
ПОЕХАЛИ?!

Проект: Аналитика сервиса «Сберавтоподписка»
по результатам работы вебсайта за 8 мес. (май—декабрь 2021г.)

Цель: Ответить на вопросы продуктологов-маркетологов
о поведении клиентов на вебсайте сервиса

Результат исследования:

Формирование выводов:
- об активности клиентов на вебсайте,
- особенностях разных групп клиентов,
- конвертации визитов в целевые действия,
- определение лидирующих характеристик клиентов,
- установление авто с наибольшим спросом
- и результативности рекламы в соц. сетях



Результат проекта:

По выводам маркетологи скорректируют
рекламную стратегию продвижения сервиса

Методология: CRISP-DM

Data Understanding — сбор и обработка данных

Дано: две исходных выборки активности клиентов на сайте сервиса

ga_sessions.pkl

характеристики визитов посетителей

Строк > 15,7 млн.
Столбцов 11

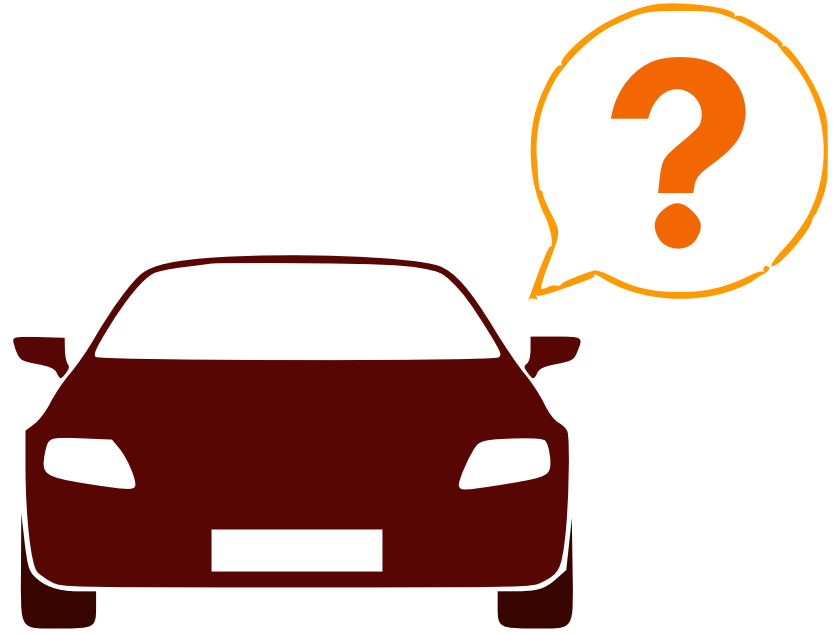
ga_hits.pkl

действия, проведенные клиентами в каждый из визитов

Строк >1, 86 млн.
Столбцов 18

Data Preparation — выбор данных

EDA — разведочный анализ данных



Особенности данных в выборках:

Все данные — строковые (кроме даты и времени)

Нет числовых данных

Часть информации зашифрована, без ключей их раскрытия

Data Preparation

EDA

Выделение целевых переменных

- Объем активности клиентов (трафик)
- Конвертация трафика в целевое действие

Основная рабочая база: **ga_sessions.pkl**

к которому добавлена информация по конвертации визита в целевое действие из файла ga_hits.pkl — признак CR

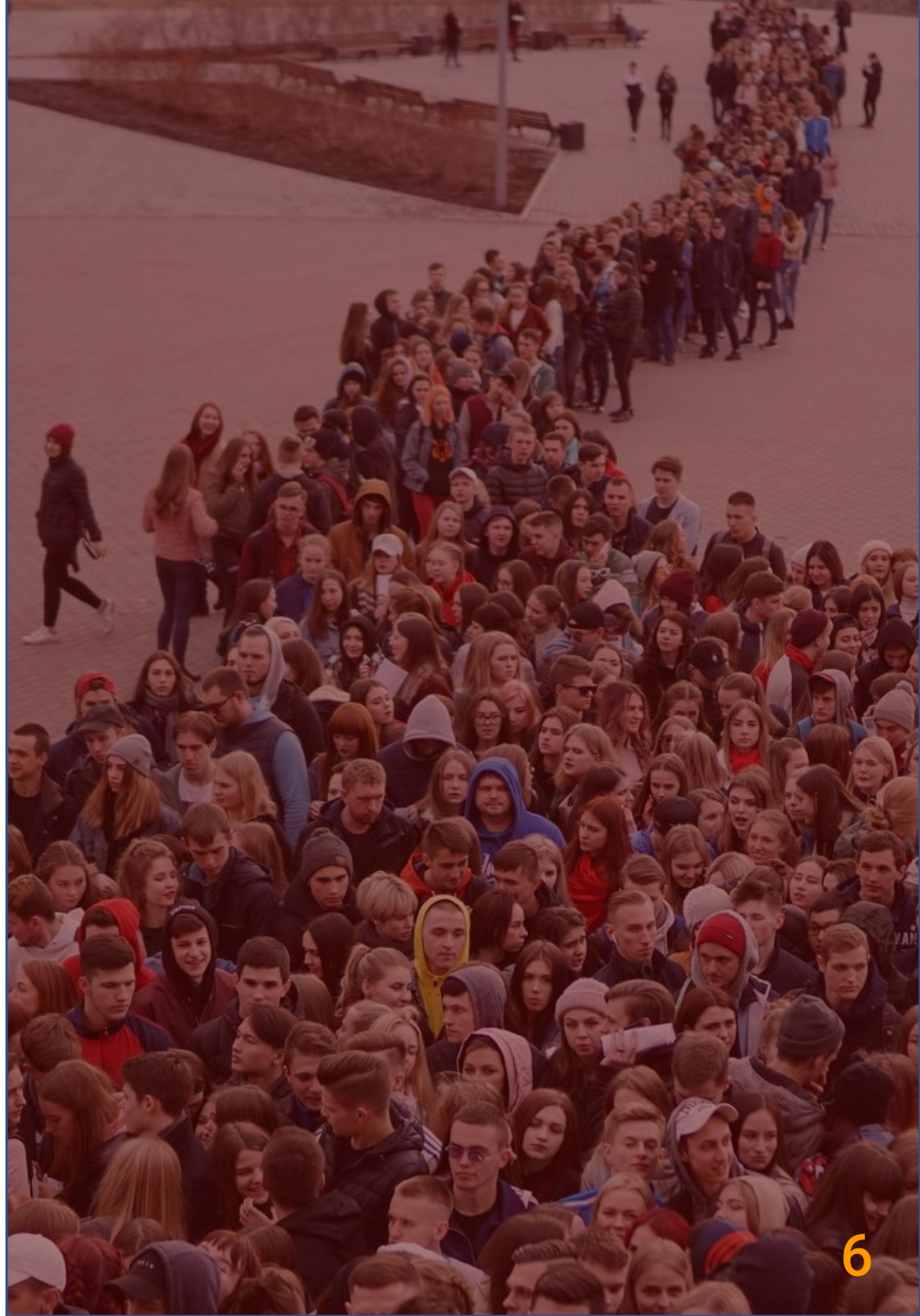
База **ga_hits.pkl** :

- для определения величины CR на визит
- для исследования марок автомобилей

Data Cleaning — чистка баз данных, подготовка

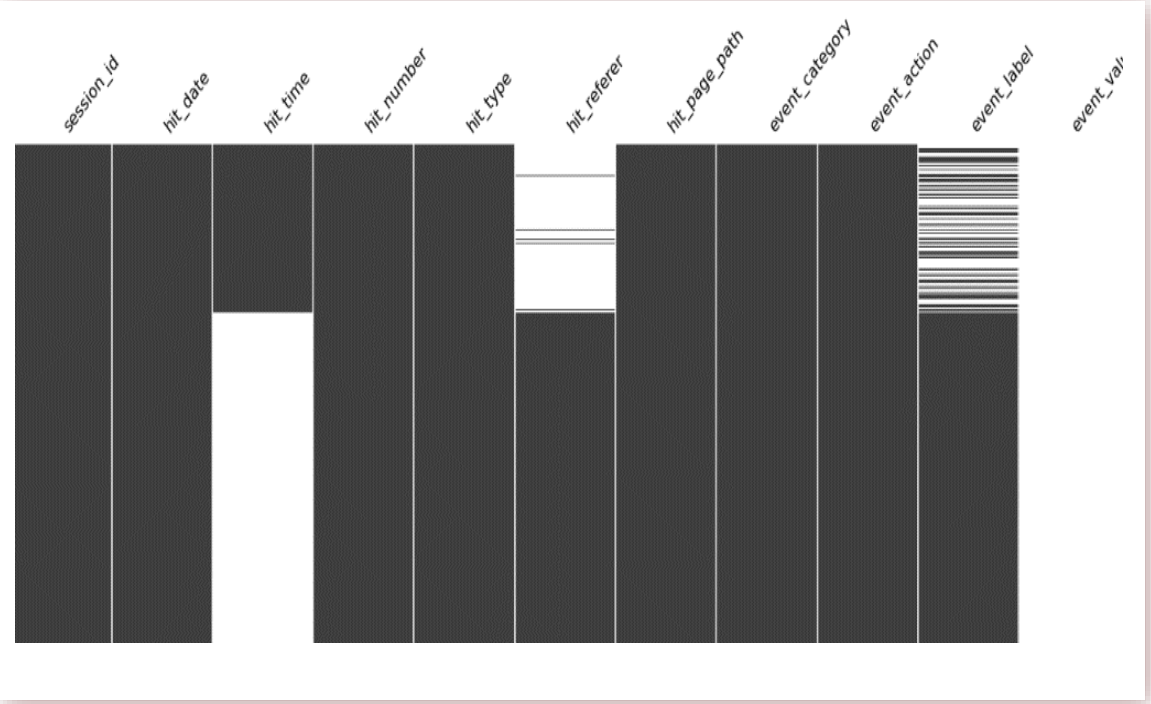
Дубликаты: дубликаты в обеих базах не установлены

Заполнение пустот: есть пробелы в исходных данных



ga_hits

Матрица ga_hits



7 столбцов из **11** заполнены на 100%

- 4 колонки не заполнены полностью — от 24 до 100% данных отсутствуют
- колонка ‘event_value’ не заполнена совсем

Сделано:

— Удалены 4 колонки, не заполненные полностью
(>24% пропущенных данных)

— Сформирован столбец CR: целевое действие

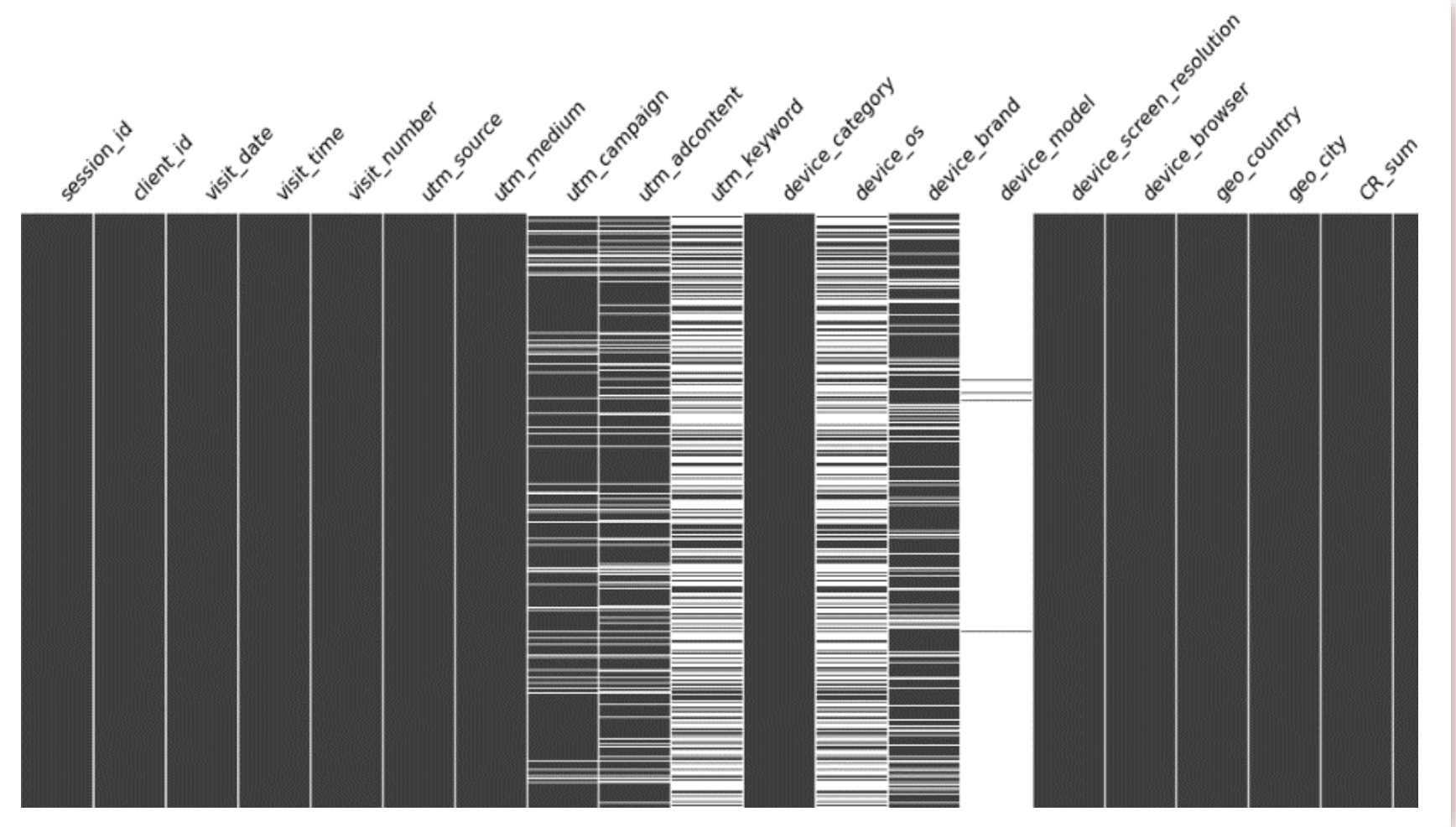
— Сформирован столбец с маркой автомобиля
(148 значений, встречаются в 3 508 732 строках, т.е. в 22,3% всей выборки)

Итоги:

все значения таблицы заполнены на 100%, без потери количества строк

ga_sessions

Матрица ga_sessions



13 столбцов из **20** заполнены на 100%

- 7 колонок не заполнены полностью — от 0,5 до 99% данных отсутствуют
- колонка 'device_model' — 99,1% данных отсутствуют
- 'utm_keyword' , 'device_os' — 59% данных отсутствуют

ga_sessions

Сделано:

- Добавлен столбец CR из ga_hits, группировка “inner”
  количество строк уменьшилось до 1 732 266 (снижение на 6,7%)
- Удалены 3 колонки, незаполненные более 58%
- Отработаны 4 колонки с отсутствующими данными и 5 колонок с нулевыми значениями (от 0,01% до 16%)

Особенности:

- Замены отсутствующих и нулевых значений проведены на моды
- 2 столбца – на моды столбцов, 6 столбцов на моды, полученные при группировке
- В городах установлены цифровые значения — также проведена замена на моды

Итого:

- Проведено заполнение 8 колонок данными
- Все значения таблицы заполнены на 100%, без потери количества строк
- Дальнейшие “докрутки” таблиц нецелесообразны с точки зрения поставленных задач

ga_sessions

Типизация данных:

- Преобразованы столбцы даты и времени визита: объединены и приведен тип данных в соответствие
- Преобразованы категориальные столбцы - 'utm_medium' и 'device_category'
- Сформированы дополнительные столбцы, необходимые для дальнейшего анализа ('CR_result' , 'org_traffic', 'advertising_social_NW', 'city_of_presence')

Удаление ненужных атрибутов

- удалены промежуточные столбцы аналитики

Результат:

- Базы готовы для дальнейшей аналитики в полном объеме

Первичный анализ —

- Первые выводы исследования – достаточно низкая конвертация в действие визитов на сайт
- То есть интерес выше, чем результат, с этим надо работать

01 Проверка гипотезы

Органический трафик больше платного с точки зрения конвертации визита в целевое событие (CR)

Проверена явная информация —

— данные не имеют нормального распределения (проверили по тесту Шапиро),

— выборки не имеют зависимости друг от друга (проверили по тесту Левена),

— поэтому для проверки гипотезы выбираем **критерий Манна-Уитни**

Подтверждена нулевая гипотеза с pvalue=0.0 —

органический трафик имеет большую конверсию в CR по отношению к платному трафику

01

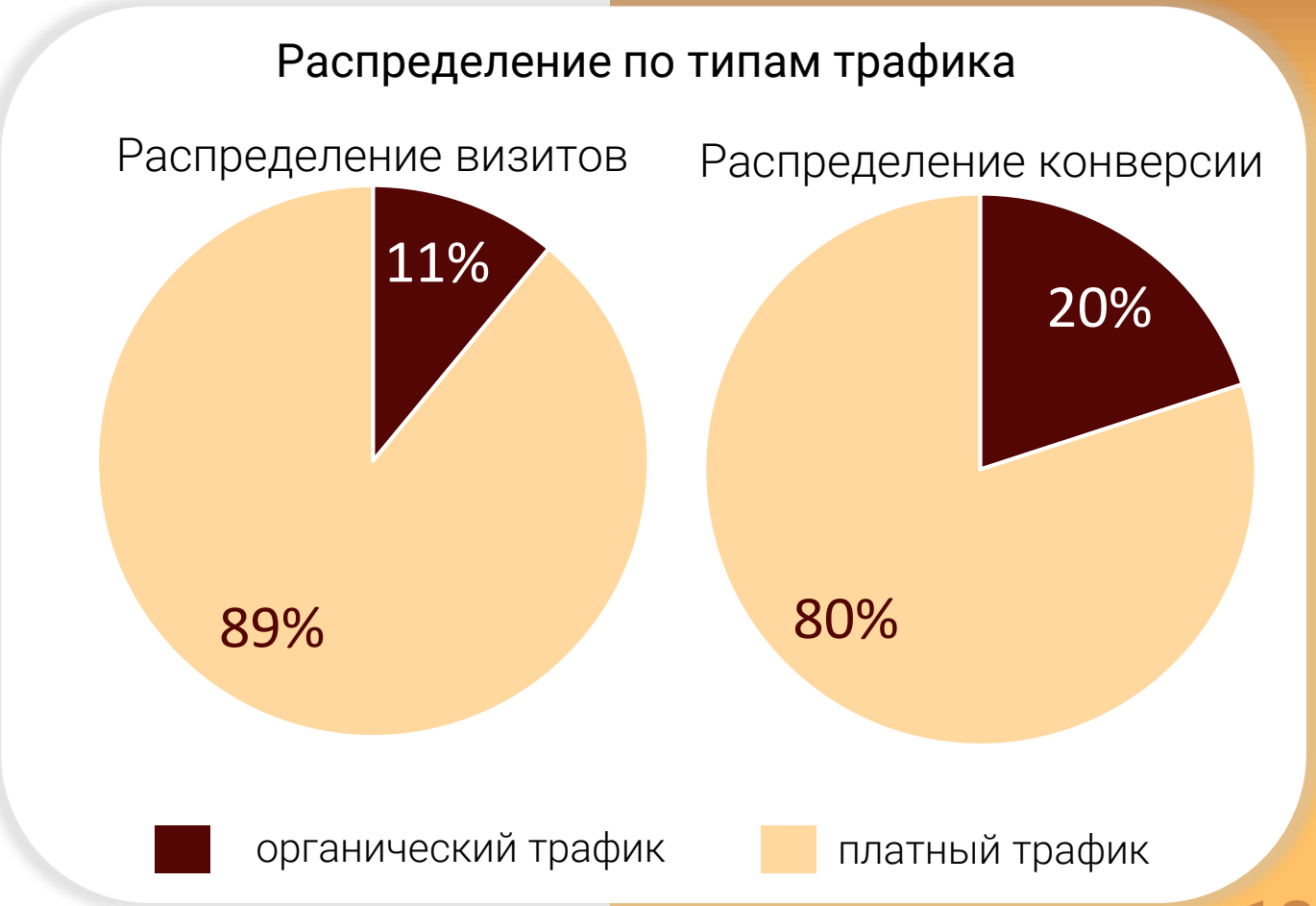
Вывод:

Органический трафик имеет большую конверсию в CR по отношению к платному трафику

Данный вывод логичен, т.к. клиент, имеющий намерение арендовать автомобиль будет в первую очередь самостоятельно заниматься поиском источников информации, в отличие от случайного приобретения по рекламной ссылке и подтверждается результатами Data Visualisation

Результат конверсии в органическом трафике практически в два раза выше, чем в платном

Это свидетельствует о большей эффективности данного типа рекламы



02 Проверка гипотезы

Трафик с мобильных устройств меньше трафика с десктопных устройств с точки зрения CR

Проверена явная информация —

— данные не имеют нормального распределения (проверили по тесту Шапиро),

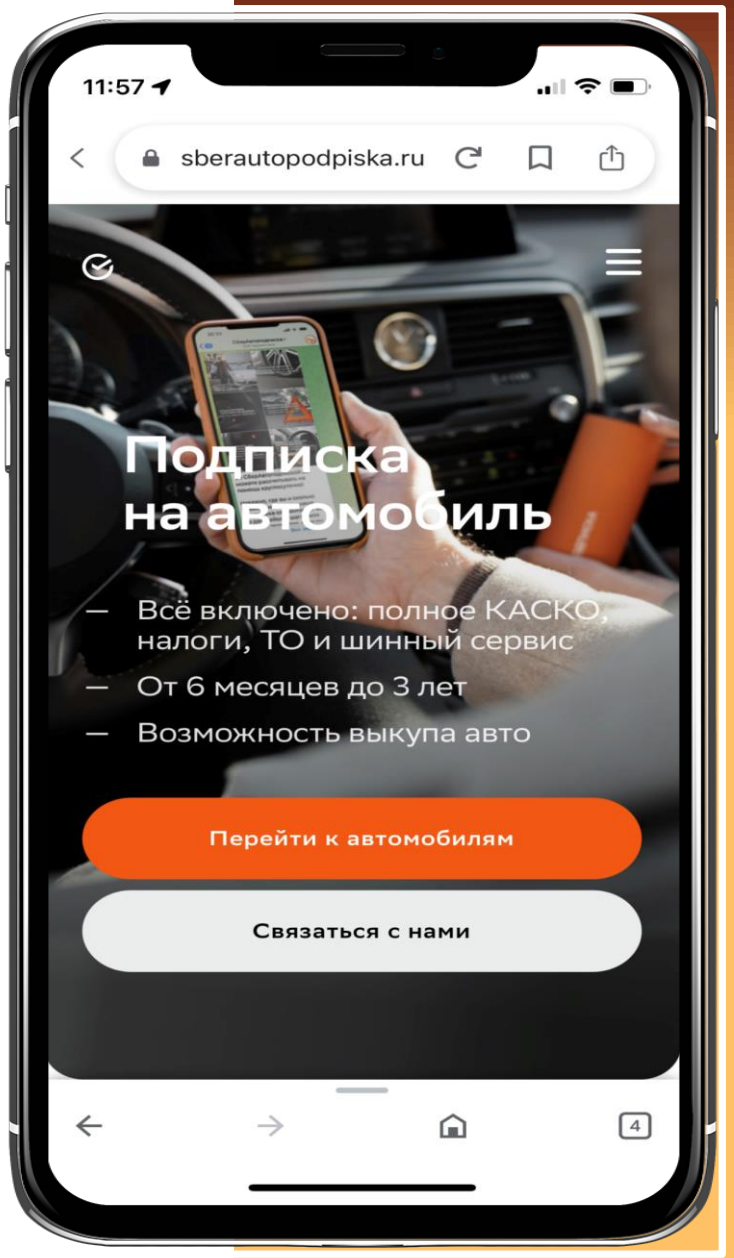
— выборки не имеют зависимости друг от друга (проверили по тесту Левена),

— поэтому для проверки гипотезы выбираем **критерий Манна-Уитни**

Подтверждена нулевая гипотеза с pvalue= 6-53 —

трафик с мобильных устройств меньше конвертируется в CR по отношению к трафику с десктопных устройств

Данный вывод логичен, т.к. сделка существенная и клиенту комфортнее ее рассматривать с помощью десктопного устройства



02 Вывод:

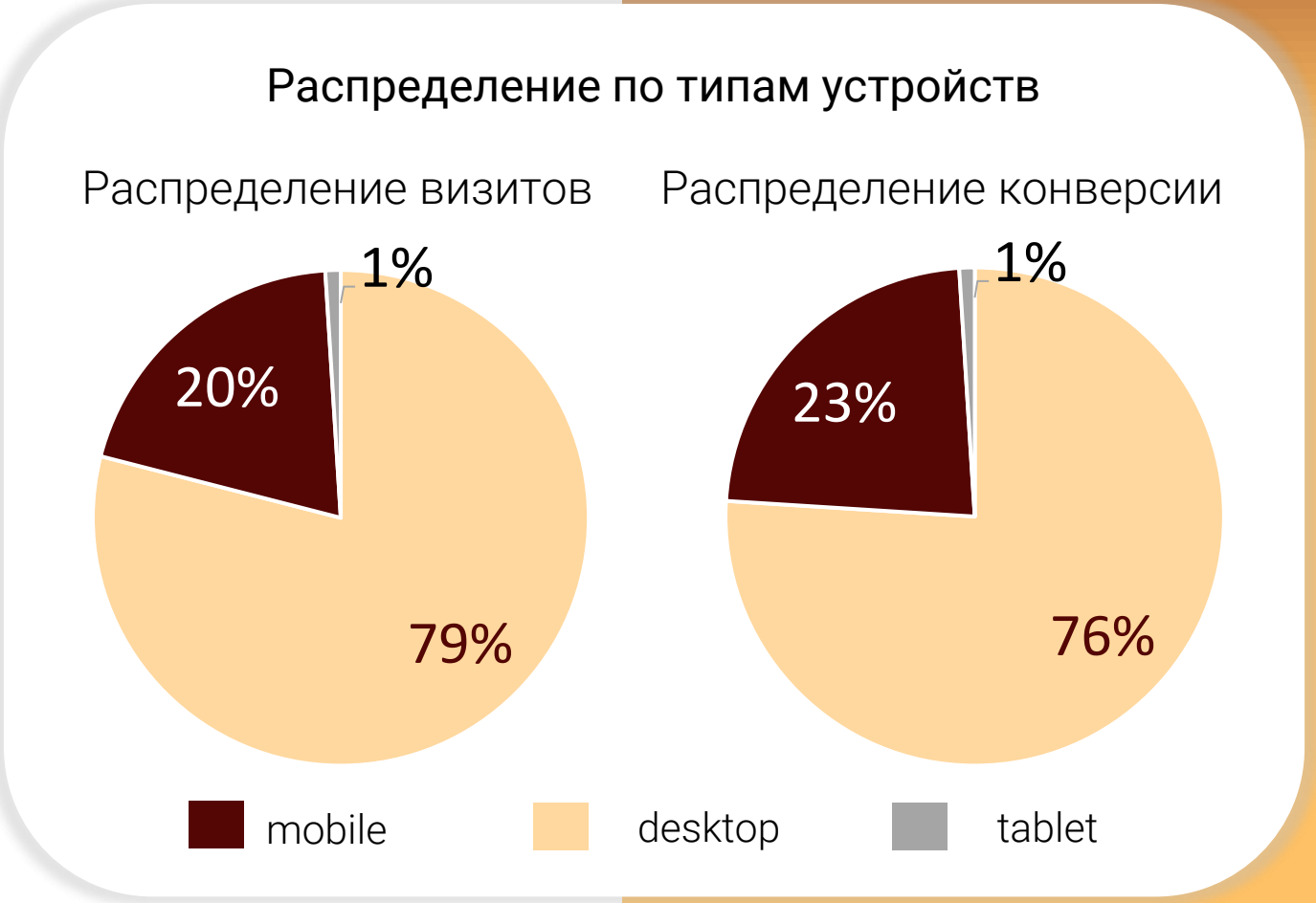
трафик с мобильных устройств меньше конвертируется в CR по отношению к трафику с десктопных устройств

Следует учитывать и результаты по Data Visualisation, из которых видно, что общее количество визитов и конверсий с мобильных устройств в три раза выше, чем с десктопных

Основное количество визитов было совершено с мобильных устройств - 79%, основное количество визитов с конверсией в целевое действие также было совершено с мобильных устройств

Процент практически аналогичен распределению визитов

Разница в конверсиях не существенна



03 Проверка гипотезы

Трафик из городов присутствия (Москва и Санкт-Петербург) больше трафика из иных регионов с точки зрения CR

Проверена явная информация —

— данные не имеют нормального распределения (проверили по тесту Шапиро),

— выборки не имеют зависимости друг от друга (проверили по тесту Левена),

— поэтому для проверки гипотезы выбираем **критерий Манна-Уитни**

Подтверждена нулевая гипотеза с pvalue= 0.0028 < 5% —

трафик из городов присутствия больше конвертируется в CR по отношению к трафику из других городов

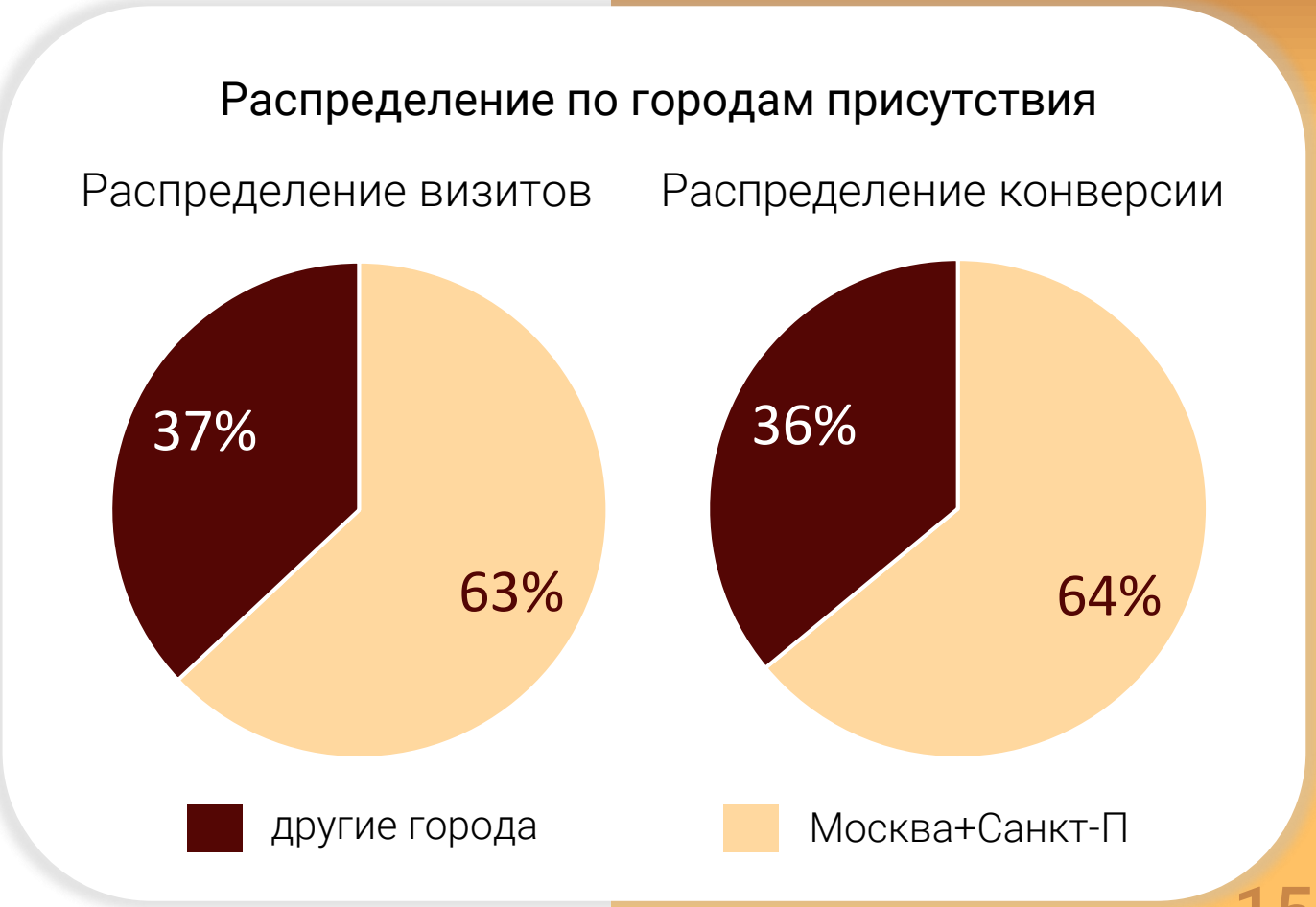
03 Вывод:

трафик из городов присутствия (Москва и Санкт-Петербург) больше трафика из иных регионов с точки зрения CR

Данный вывод логичен, т.к. и по результатам по Data Visualisation видно, что общее количество визитов и конверсий из городов присутствия значительно выше, чем из других городов, однако также заметно, что и разница в конверсиях незначительная

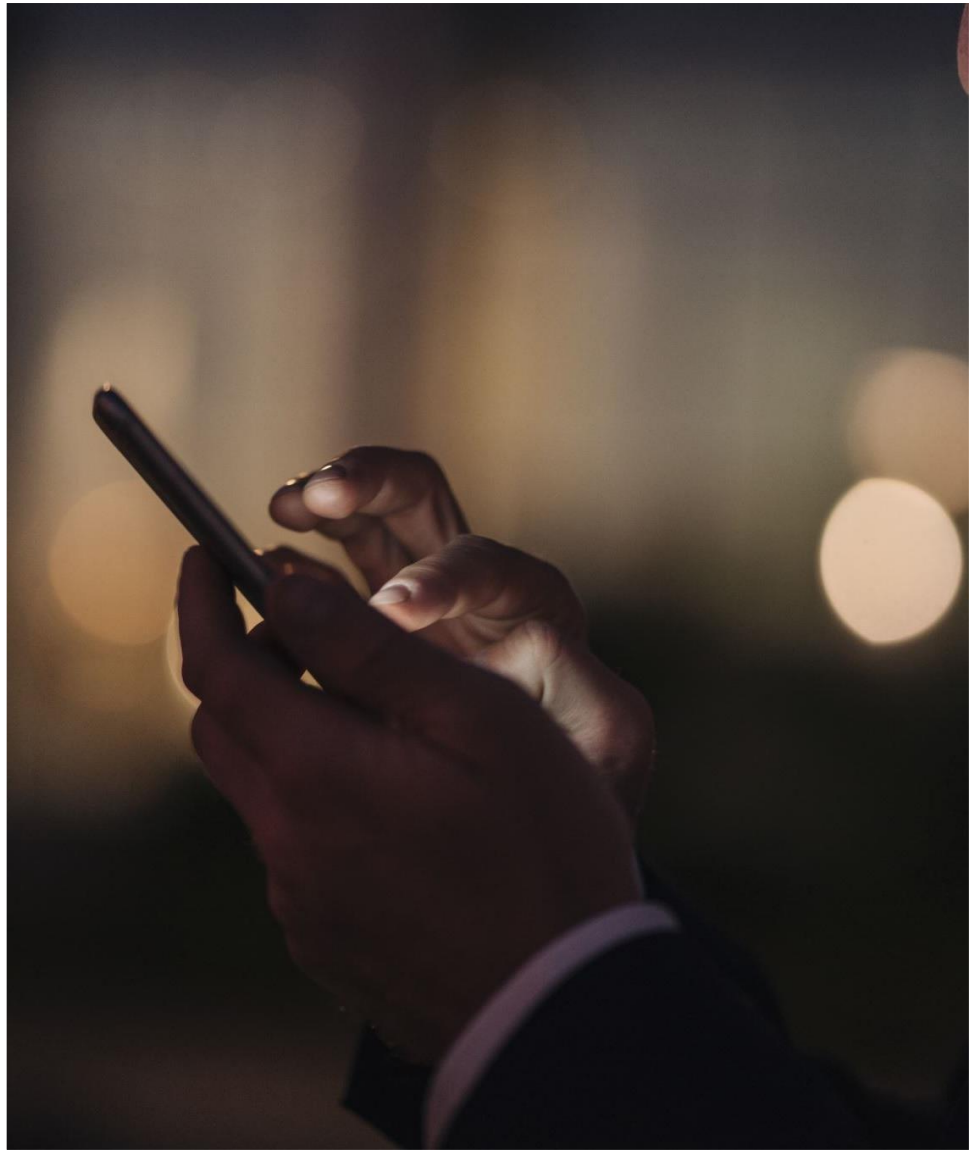
Результат конверсии по городам присутствия практически не отличается от результата по другим городам,

что свидетельствует о незначительности результата конверсии от принадлежности к городам присутствия



04 Проверка гипотезы

Реклама в соц. сетях дает больший результат, чем реклама в иных средствах



Проверена явная информация —

— данные не имеют нормального распределения (проверили по тесту Шапиро),

— выборки не имеют зависимости друг от друга (проверили по тесту Левена),

— поэтому для проверки гипотезы выбираем **критерий Манна-Уитни**

НЕ подтверждена нулевая гипотеза с pvalue= 1 —

реклама в соц. сетях не дает больше конверсии CR в целевые события, чем иные виды рекламы

04 Вывод:

Реклама в соц. сетях дает меньше конверсии CR в целевые события, чем иные виды рекламы

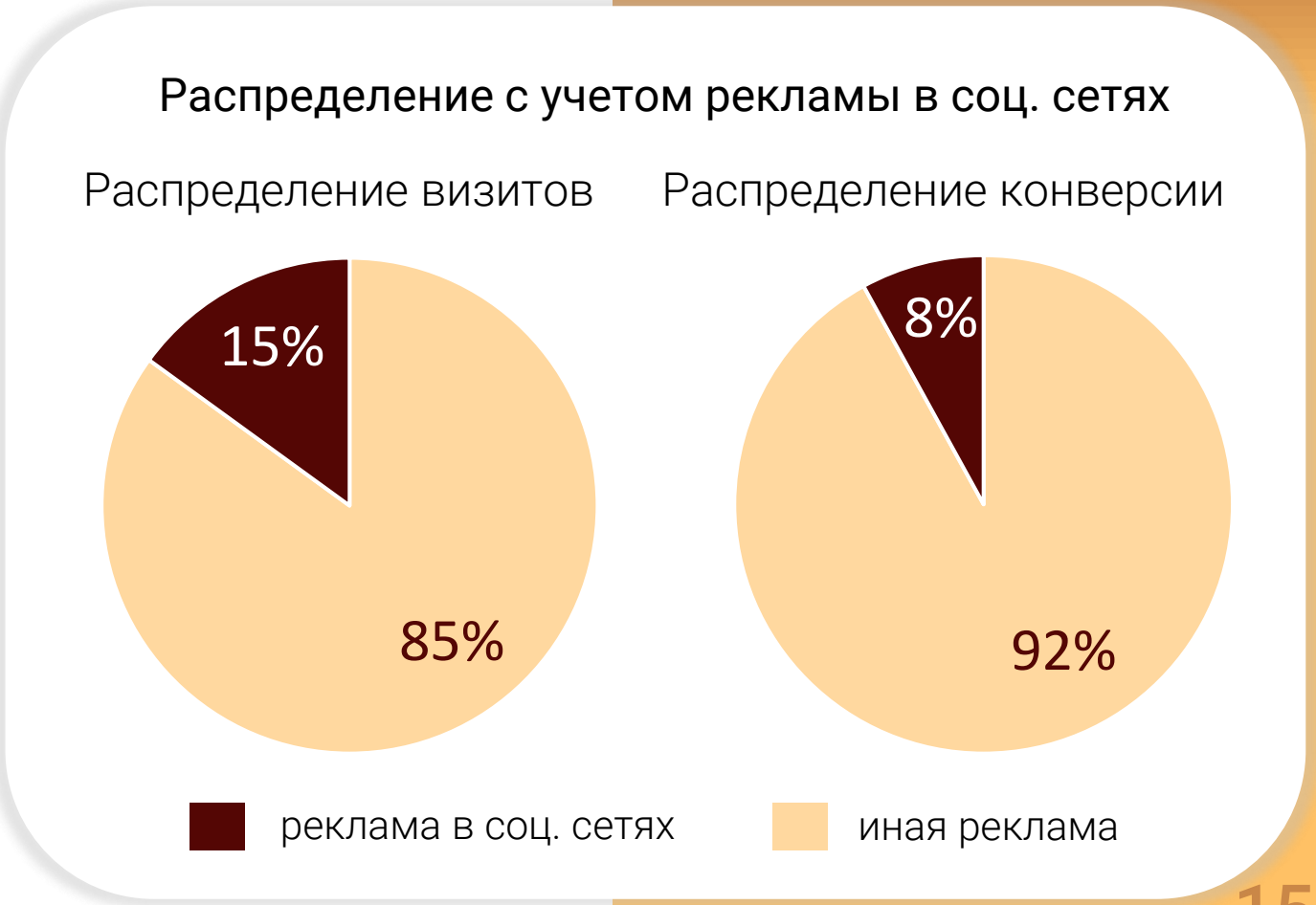
Итоговый вывод —

нет необходимости увеличивать свое присутствие в соц. сетях и давать больше рекламы в ней.

Данный вывод логичен и подтверждается, в т.ч. результатами по Data Visualisation

Результат конверсии по рекламе в соц.сетях почти в два раза ниже, чем по другим видам рекламы,

что свидетельствует о нецелесообразности развития данного способа рекламы продукта (услуги)



Рассмотрим лидеров!

05.1

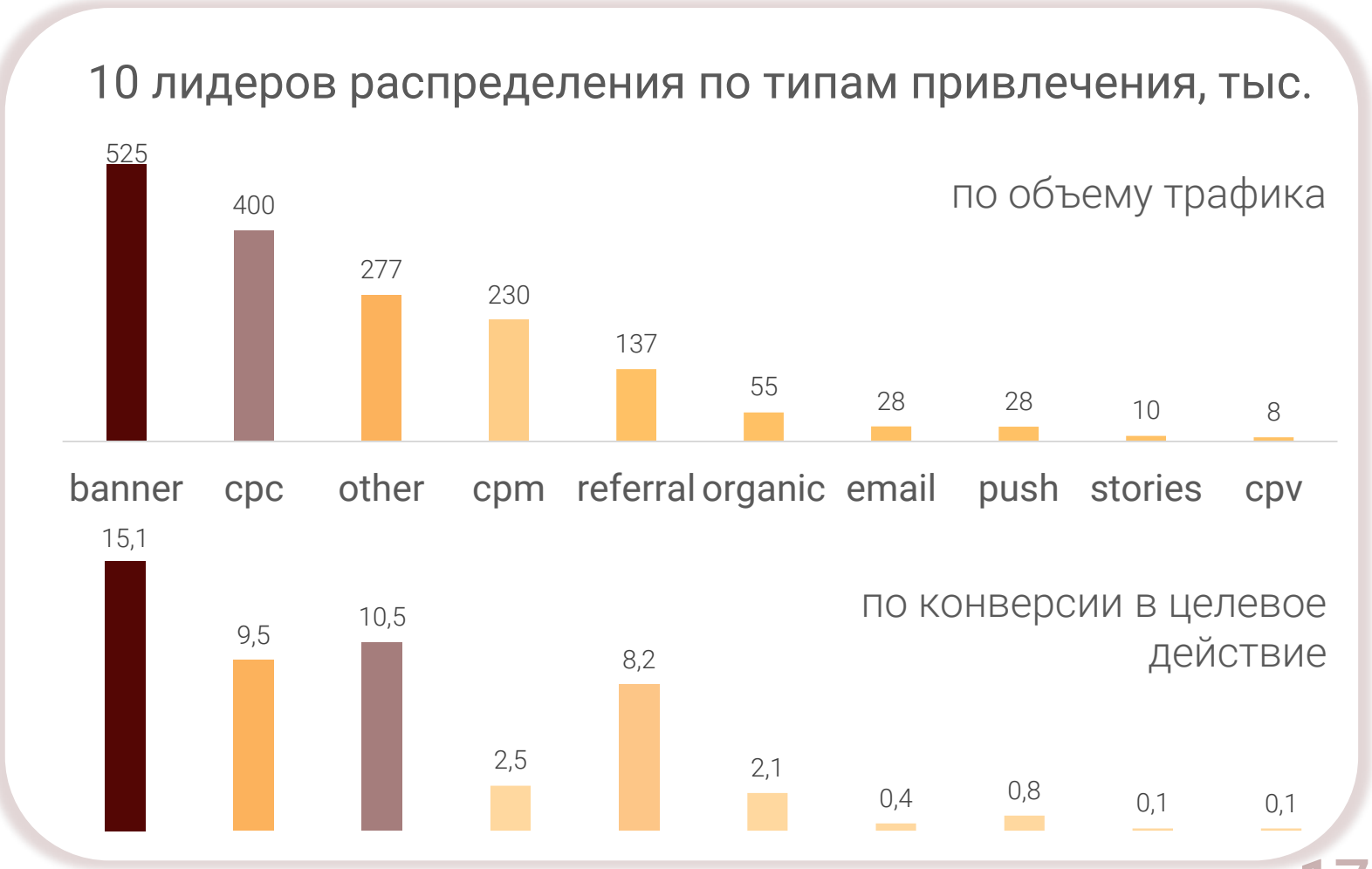
Выделение лучших

Из каких ИСТОЧНИКОВ/кампаний/устройств/локаций идет самый целевой трафик — и с точки зрения объема, и с точки зрения CR

Баннер —

привлечение основного количества визитов

Распределение по типам привлечения в части 10 лидеров практически совпадают по распределению по визитам с распределением с результатов конверсии в целевое действие



05.1

Выделение лучших

Из каких источников/КАМПАНИЙ/устройств/локаций идет самый целевой трафик — и с точки зрения объема, и с точки зрения CR

График рекламных кампаний показывает четверку явных лидеров. Установить явные признаки кампании не удалось

05.1

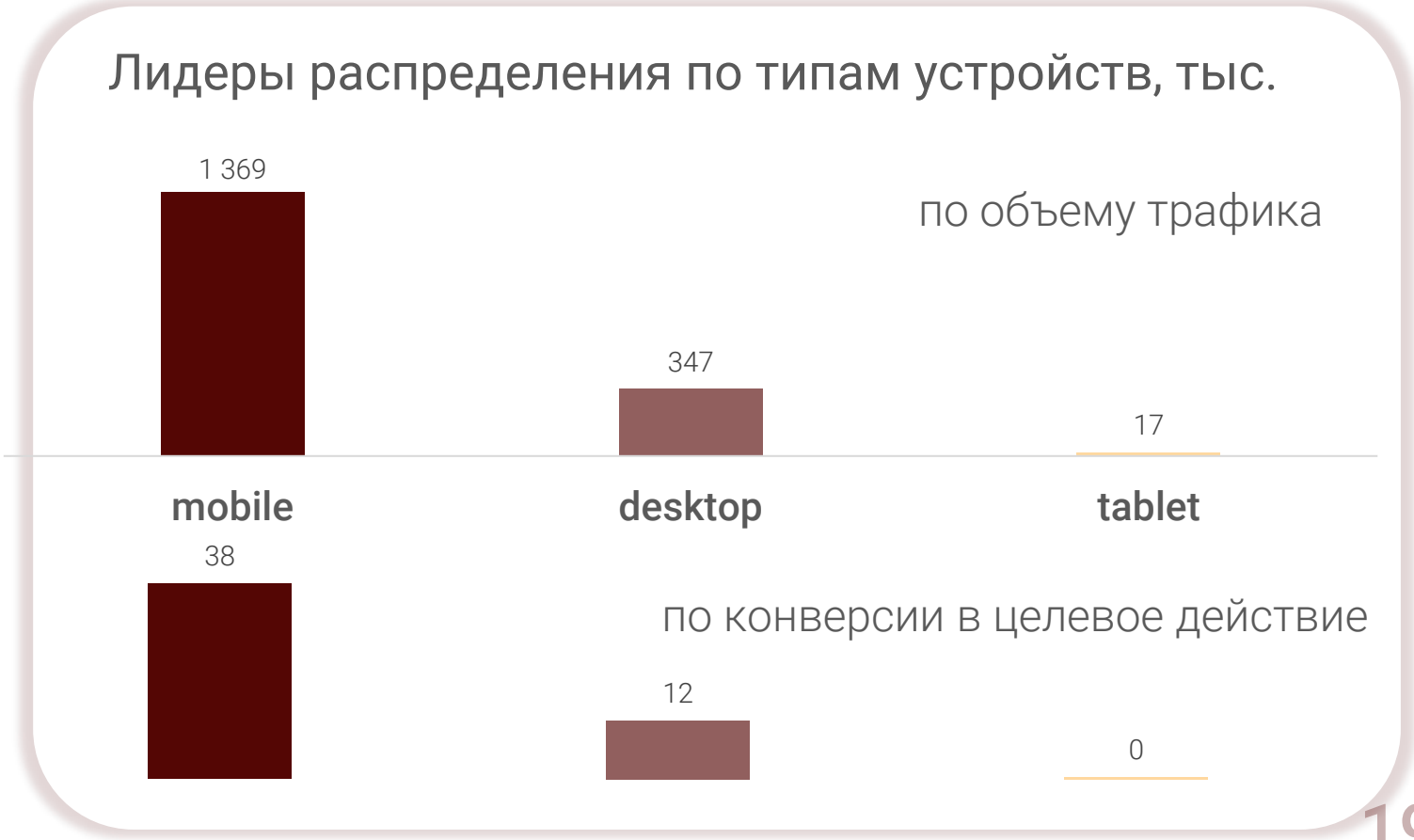
Выделение лучших

Из каких источников/кампаний/**УСТРОЙСТВ**/локаций идет самый целевой трафик — и с точки зрения объема, и с точки зрения CR

79% визитов было совершено с мобильных устройств

Основное количество визитов с конверсией в целевое действие также было совершено с мобильных устройств

Процент практически аналогичен распределению визитов

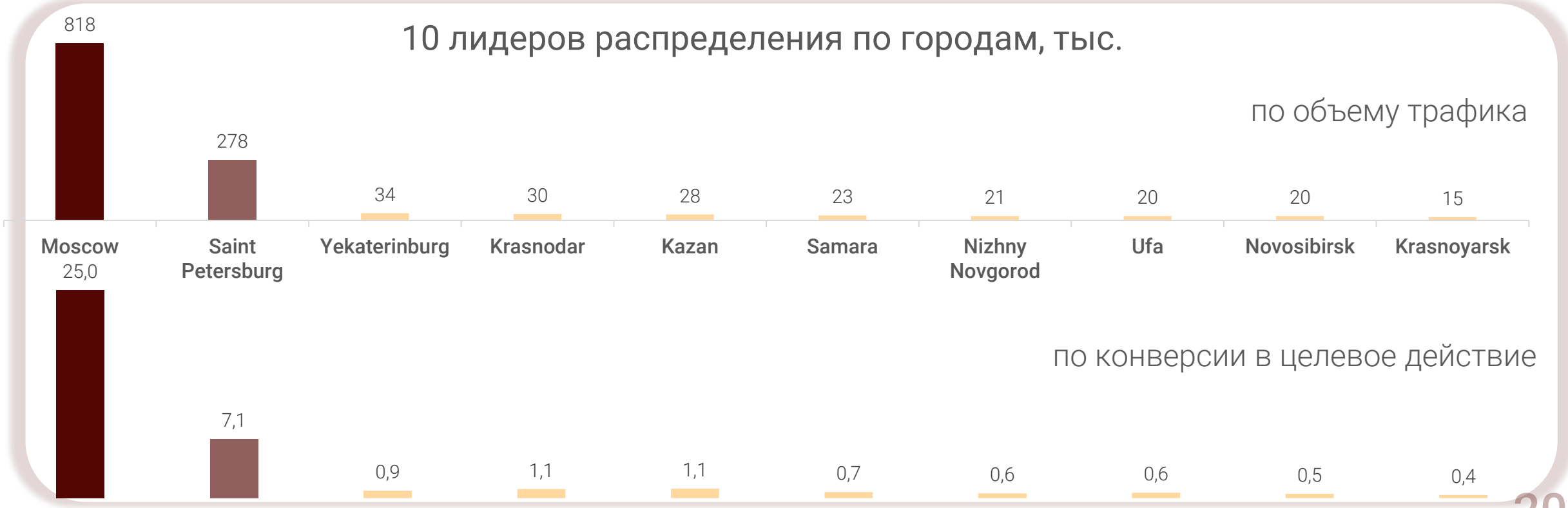


05.1

Выделение лучших

Из каких источников/кампаний/устройств/ЛОКАЦИЙ идет самый целевой трафик — и с точки зрения объема, и с точки зрения CR

«Города-миллионники» — лидеры визитов и конверсии их в целевое действие



05.2

Выделение лучших

Какие авто пользуются наибольшим спросом и какие авто имеют лучший показатели CR

Skoda, Lada, Volkswagen polo — авто эконом класса имеют наибольший спрос

В пятерку лидеров также попал mercedes-benz e-klasse, т.е. авто среднего класса стоимости.
Среди лидеров нет ни одной марки авто высокого класса стоимости

Подведем итоги?

06.1 Итоговые выводы проекта

Бо′льшую конверсию в CR имеют:

- ✓ органический трафик по отношению к платному трафику
- ✓ трафик с десктопных устройств по отношению к трафику с мобильных
- ✓ трафик из городов присутствия больше трафика из иных регионов
- ✓ иные виды рекламы в сравнении с рекламой в соц.сетях

Наибольший спрос у машин:

самого экономного класса – Skoda, Lada, Volkswagen polo

как в визите, так и в конверсии

Лидеры объема трафика и CR:

- среди источников – баннеры, cpc
- среди кампаний – 'LTuZkdKfxRGVceoWkVyg'
- среди устройств – мобильные устройства
- среди локаций – города присутствия

06.2 **Дополнительно:** рекомендации

Предлагается маркетологам провести более подробные исследования по темам:

1. Динамика активности клиентов

совместив с периодами проведения рекламных кампаний, при необходимости – добавить регионы, города

По результатам можно сделать выводы по особо удачным рекламным кампаниям и их отличиям по городам

2. Зависимость трафика и CR от стоимости авто

ранжировав марки авто по стоимостным группам (исследование провести по каждой ценовой категории)

По результатам можно сделать выводы о наиболее популярных авто в каждом сегменте и, возможно, пересмотреть предлагаемую типовую линейку авто.

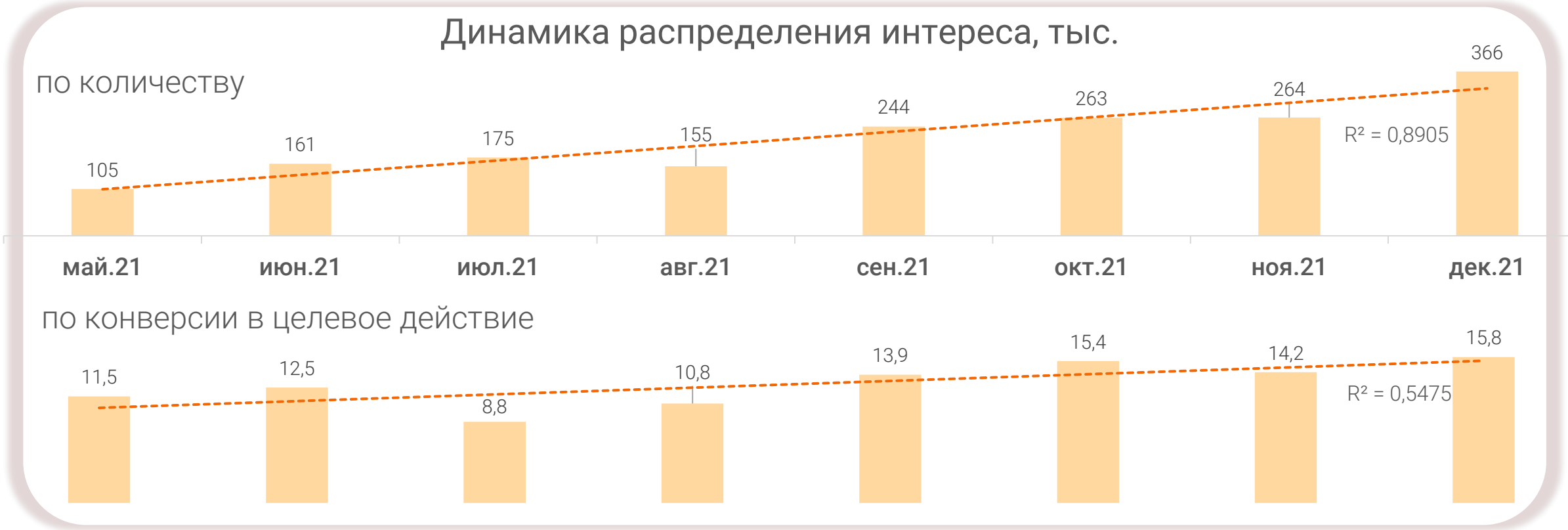
Кроме этого, это позволит более четко определить целевые аудитории каждой категории авто и уточнить рекламную стратегию сервиса

3. Особенности активных клиентов (по трафику и CR)

кто, откуда, какими устройствами пользовались и пр.

Возможно, это описание позволит сделать дополнительные выводы о характеристиках целевой аудитории

06.3 **Дополнительно:** исследование динамики активности клиентов по периодам



По динамике виден практически непрерывный тренд нарастания общего количества визитов клиентов, но при этом тренд конверсии выражен меньше. Целесообразно исследовать рекламные кампании в периоды снижения и роста как количества визитов, так и конверсии

СПАСИБО
за внимание!