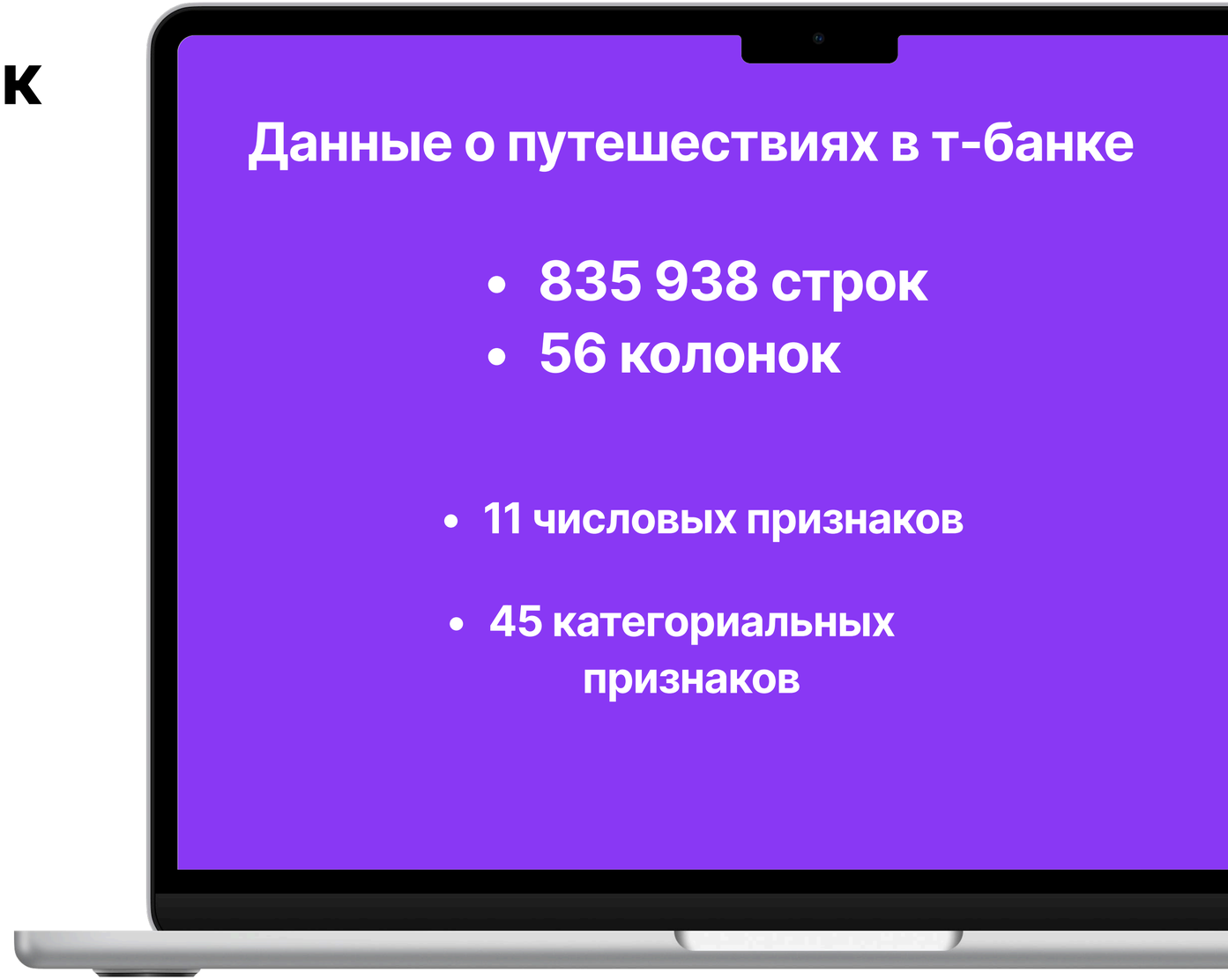
Отбор на стажировку тбанк

Структура данных

делал я



Предварительная обработка данных

- Удаление колонок, в которых слишком много пропусков
- Удаление колонок с идентификаторами, которые не несут информации
- Заполнение пропусков, где возможно
- Удаление строк, в которых невозможно определить тип поездки AIR/HOT (58415 строк - 6.99%)

Осталось строк: 777523, 93.01%

Осталось 46 столбцов, удаленные столбцы:

- bad_email_address_flg
- free_cancel_booking_dttm
- call_contact_1m_flg
- call_contact_3m_flg
- call_contact_6m_flg
- cancel_dttm
- good_email_address_flg
- last_email_send_dt
- account_rk
- order_rk

Предварительная обработка данных

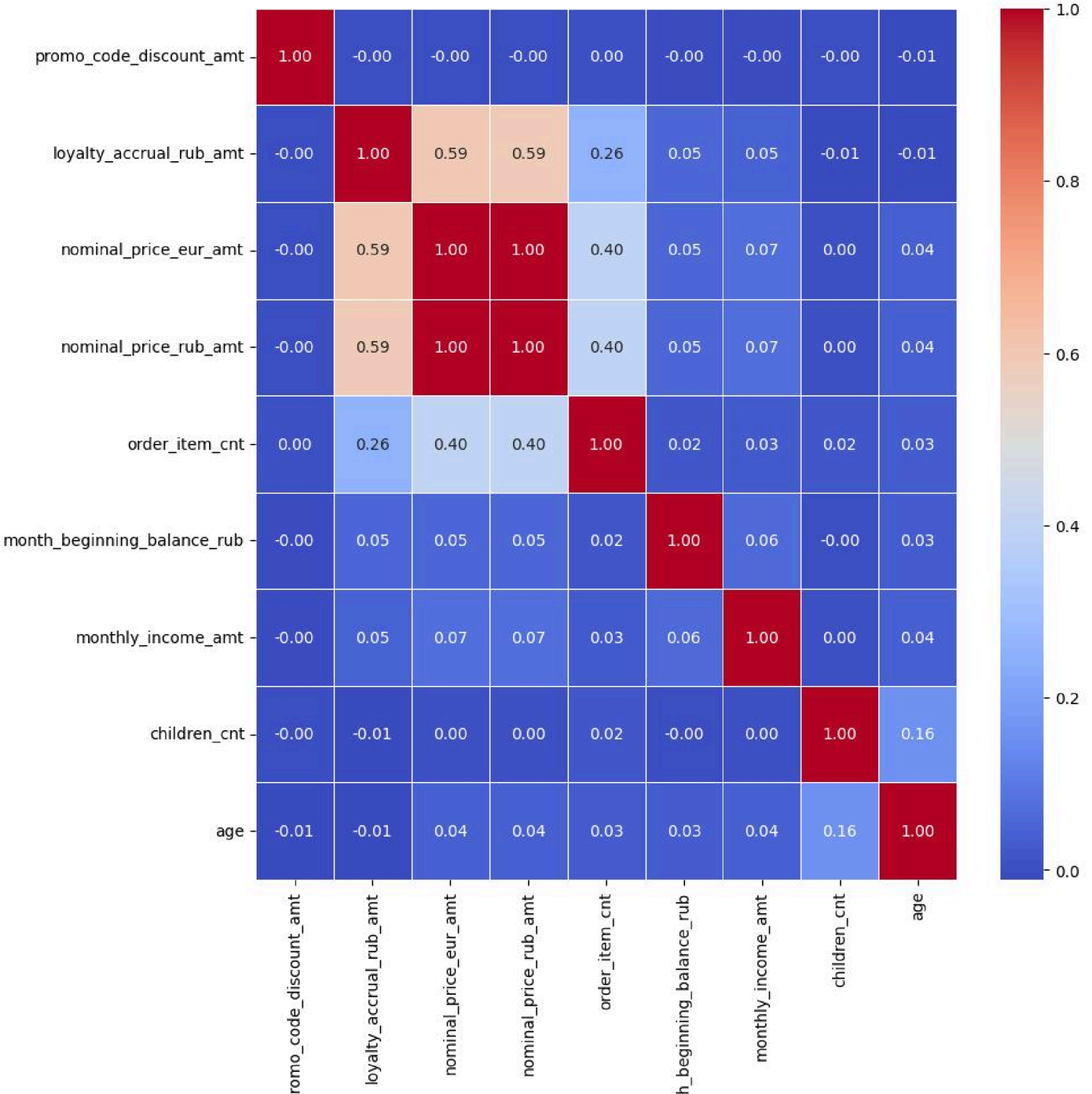
- Заполняем пропуски по региону проживания городом, при отсутствии удаляем
- Создаем отдельные классы для не указанной степени образования и семейного статуса.
- Удаляем выбросы
- Отсеиваем по программе лояльности.
- Отсеиваем по возрасту.
- Заполняем нулями отсутствие бонусов тем, кто действительно не пользовался программой лояльности.
- Удаляем пропуски по важным колонкам.
- Удаляем колонки, в которых очень много пропусков.
- Удаляем незадействованные колонки
- Удаляем строки, в которых невозможно определить AIR/HOT

Итого осталось:
543 989 строк - 65.07 % данных

Предварительный анализ

Выявление линейных зависимостей в данных Построение матрицы корреляций Пирсона

Главная зависимость - начисление баллов лояльности от цены,коэффициент корреляции Пирсона равен 0.59.



Предварительный анализ

Обучим и проверим модель на разных категориях данных на поиск зависимости кешбека от цены
Уберем строки, в которых не получены баллы лояльности

На покупках авиабилетов:

P-value: 0
R-squared: 0.841
F-statistic: 1.685e+06
Коэф. регрессии: 4.96

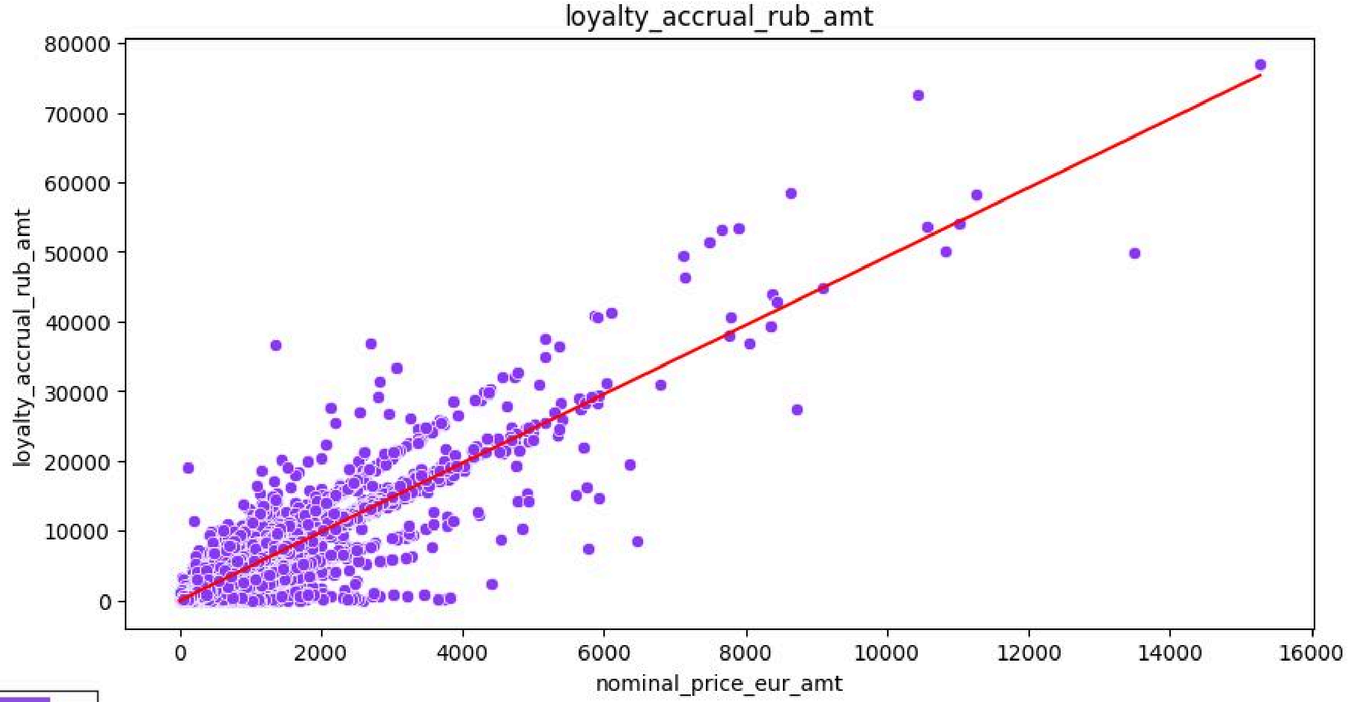
На брони отелей:

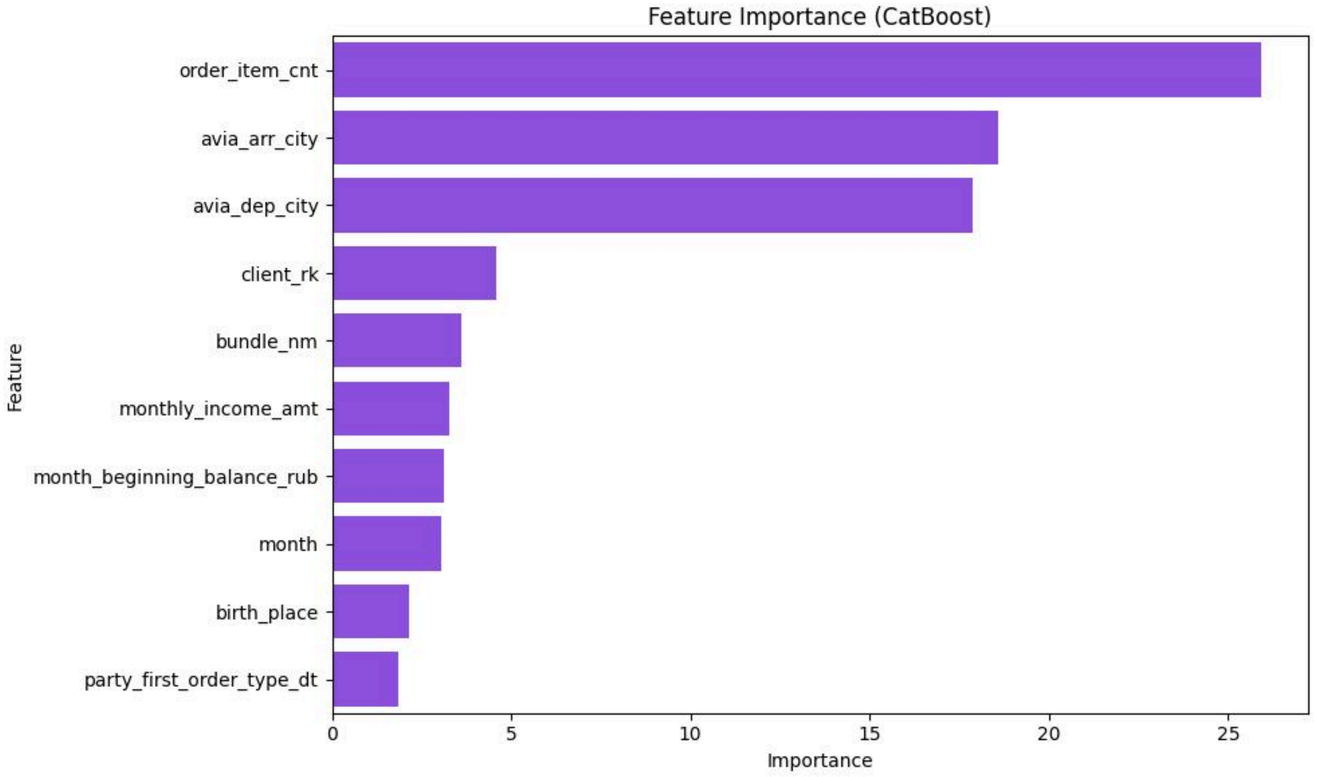
P-value: 0
R-squared: 0.159
F-statistic: 1.685e+06
Коэф. регрессии: 1.11

Как видно из метрик, в категории брони отелей наша модель не стат значима вообще, однако покупки авиабилетов модель очень хорошо описывает.
Отбросим категорию отелей и будем подробно рассматривать авиабилеты.

Предварительный анализ

График зависимости суммы кешбека от стоимости полета





Я обучил CatBoost для предсказания стоимости билетов, и самые важные признаки для модели это **локация** откуда летят и которую отправляются, а также **количество билетов**.

Можно заметить, что есть клиенты с большим количеством заказов и из их истории также можно предсказать стоимость их полета

Исследовательский вопрос

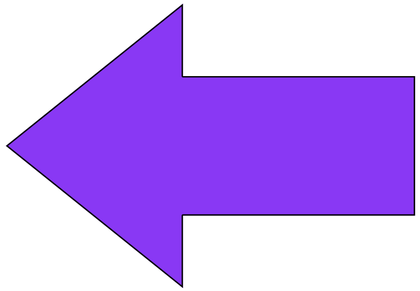
Необходимо рассмотреть всех покупателей и понять:

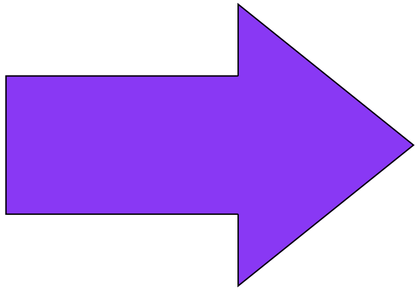
Какие покупатели чаще пользуются сервисом?

Какие покупатели тратят больше средств?

Гипотеза

Люди, состоящие в браке или имеющие детей чаще пользуются сервисом для повторных покупок.

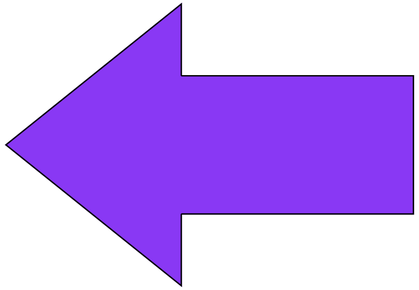


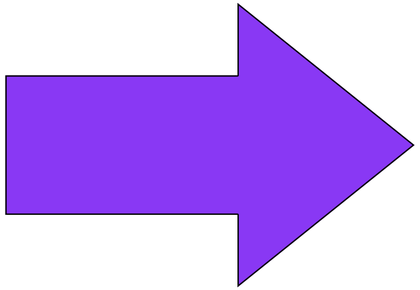


Люди часто летающие заграницу или на большие расстояния чаще пользуются сервисом

Механизм гипотезы

Люди, состоящие в браке или имеющие детей тратят больше средств, чтобы купить билеты для всей семьи ⇒ получают больше баллов лояльности.





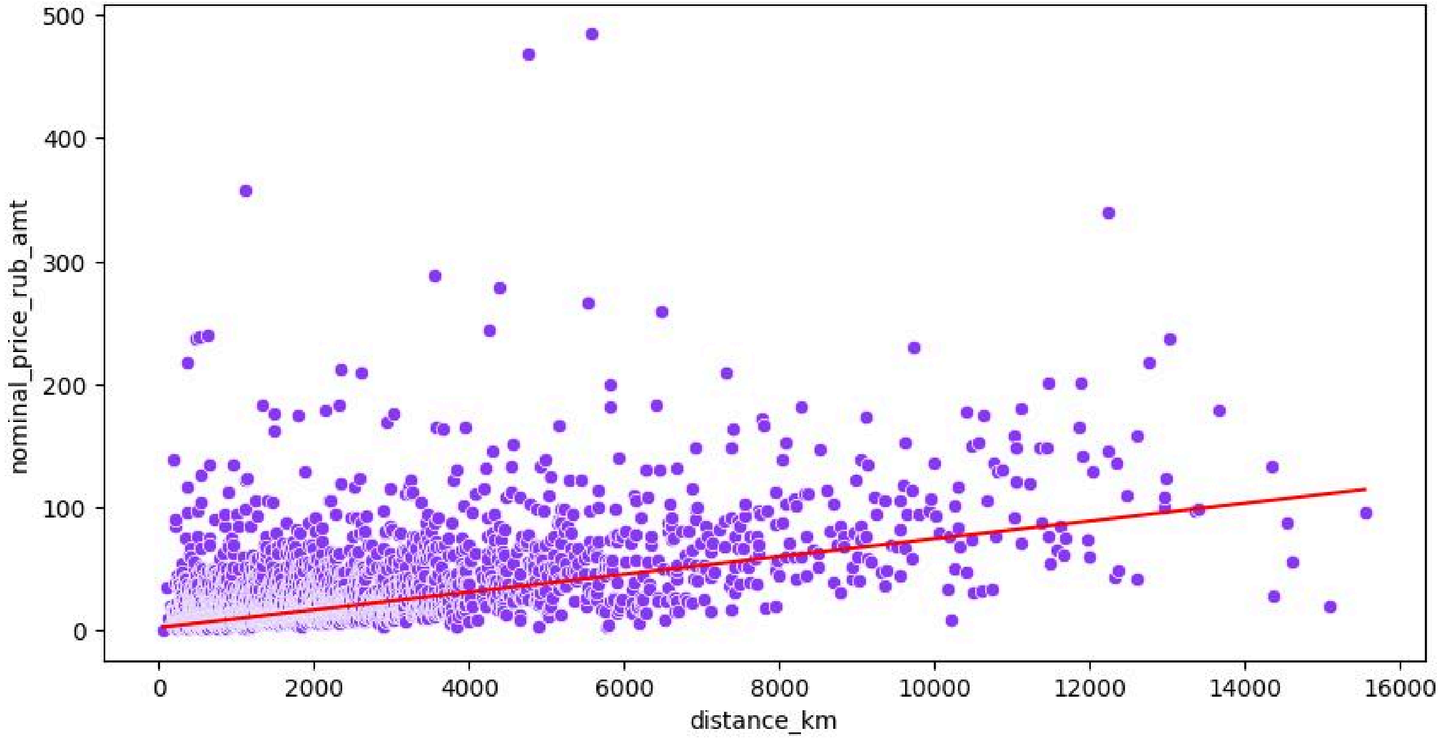
С дальних и заграничных перелетов люди получают больше баллов лояльности из-за стоимости ⇒ с большей вероятностью снова воспользуются сервисом.

Анализ данных

Анализ зависимости цены(тыс. руб) от расстояния

Были получены следующие статистики:

Корреляция Пирсона - 0.41
R-squared: 0.170
p-value: 0



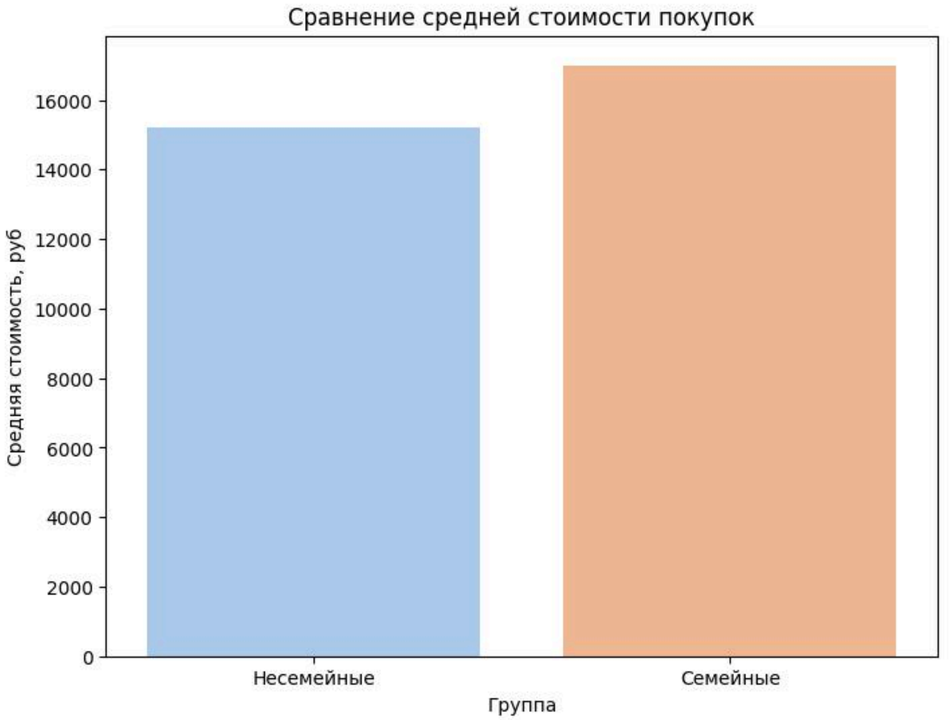
Вывод:

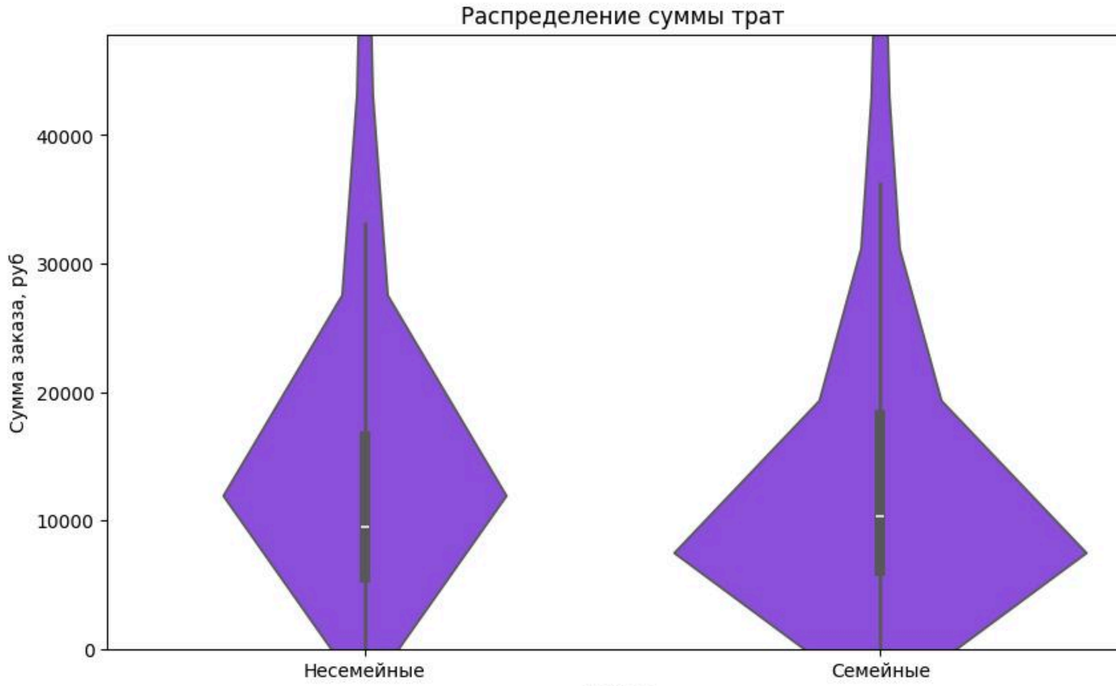
Стоимость билетов зависит от расстояния,
но сильной зависимости нет.
Скорее всего цена зависит от города из-за
внешних факторов, таких как сезонность,
объем тур потока в конкретном городе,
спрос, класс полета и тд.

Анализ данных

Семейный	Средняя стоимость билетов	Медиана
Да	16 936 руб	10329
Нет	15 223 руб	9506

Семейные пары действительно в среднем тратят больше средств, и дорогих покупок у них больше ⇒ начисления баллов лояльности больше, и они могут тратить их на повторные покупки





Стат тесты

T-test	U-test	Краскела-Уоллиса
p-value=0.000	p-value = 0.000	p-value = 0.0000
Cohen’s d: 0.07	Cliff’s delta: 0.06	Cliff’s delta: 0.017

Значимых различий между подгруппами нет, следовательно гипотеза подтверждается для всех подгрупп

Ограничения

Не указаны многие важные данные, например прежняя история покупок, род деятельности человека.
Однако не смотря на ограничения выбранную подгруппу модель хорошо описывает.

Выводы

Модель хорошо описала покупки Авиабилетов.
Для повышения прибыли можно предлагать семьям с детьми акции за повторное количество покупок или расстояний, а вследствие получения большего количества бонусов они продолжат покупать билеты в сервисе.
Не семейным людям также предлагаюстя бонусы за повторные покупки.
Также рассмотреть бонусы за длинные перелеты