



Олимпиада Dano

Анализ КАСКО

Команда 13

Структура данных

В датасете дана информация о покупке полиса КАСКО, и владельцах авто, купивших ее.

Количественные:

- Сумма премии выпущенного полиса
- Количество убытков за предыдущие периоды
- Стоимость предыдущего полиса страховой
- Сумма предполагаемого убытка по полису

Качественные:

Информация о клиенте –

- Пол
- Возраст
- Стаж вождения

Информация о заказе и авто -

- id заказа
- Тип покрытия полиса
- Модель, цена и год выпуска авто
- Регион использования авто
- Различные флаги по полису

Обработка датасета

Удаление дублей

Обработка дублей.

В дата сете есть дубли. Уникальная связка на человека - это `policy_rk+gender+age`.

Если дубли остаются, то дальше дедублицируем `expected_loss_amt = max`, если и после этого остались, то дедублицируем `min_driver_experience_years_cnt = max`

Изначально было 148476 записей, удалили 122586, осталось 25890.

Обработка датасета

Удаление выбросов

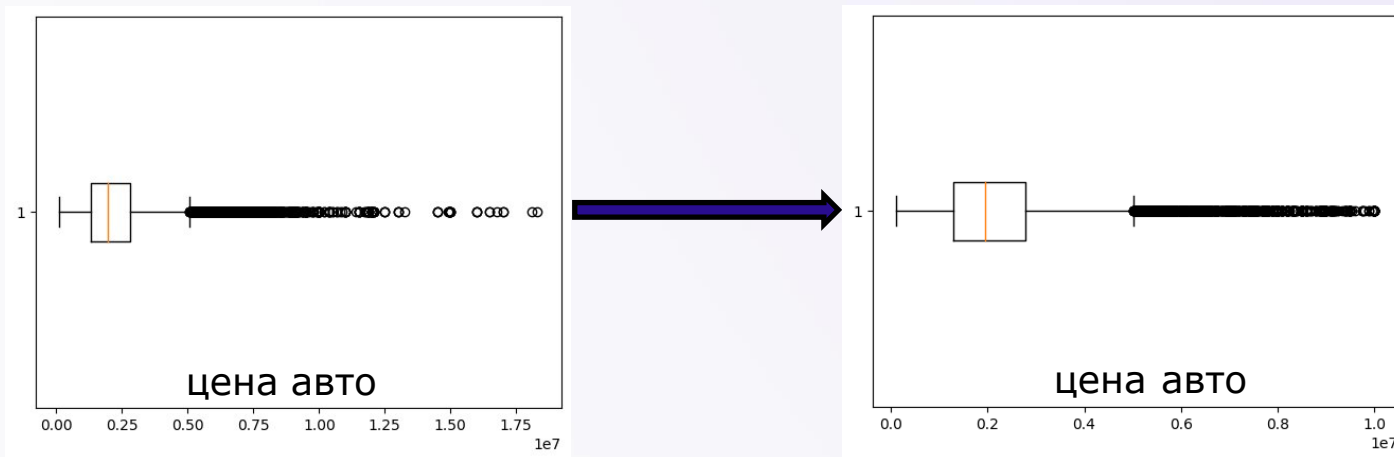
Удалили ненужный столбик `issue_flg`, так как во все всех строках значение = 1.

Убрали 57 строк, где `expected > auto_cost`, ведь это странно, что предполагаемая цена убытка больше цены авто.

Удалили 564 выброса, где `возраст - стаж - 18 < 0`, так как автоводители получают права после 18 лет.

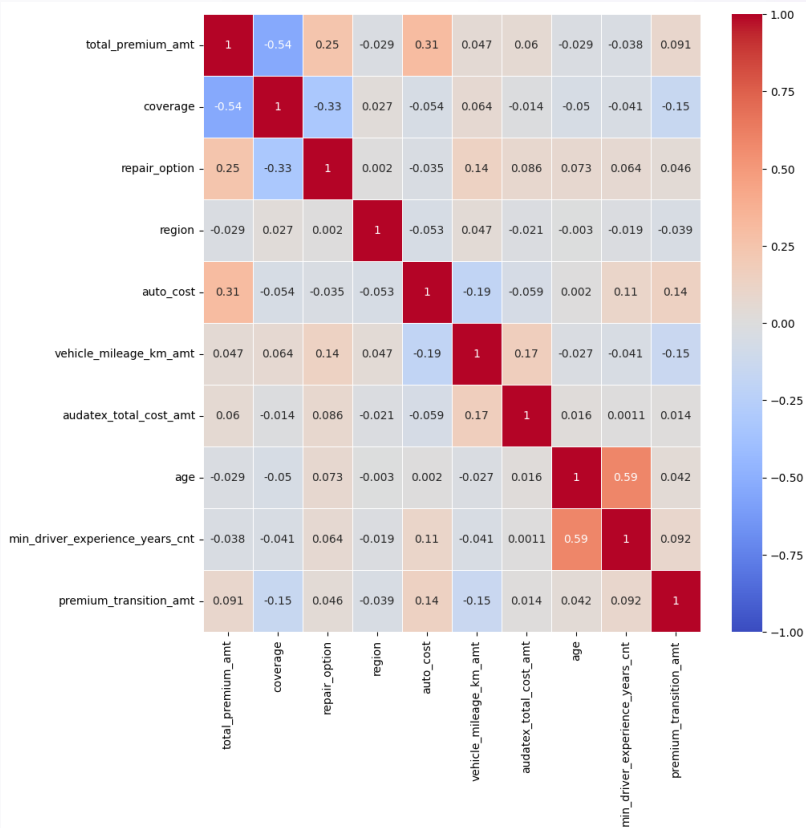
Обработка датасета

Удаление выбросов

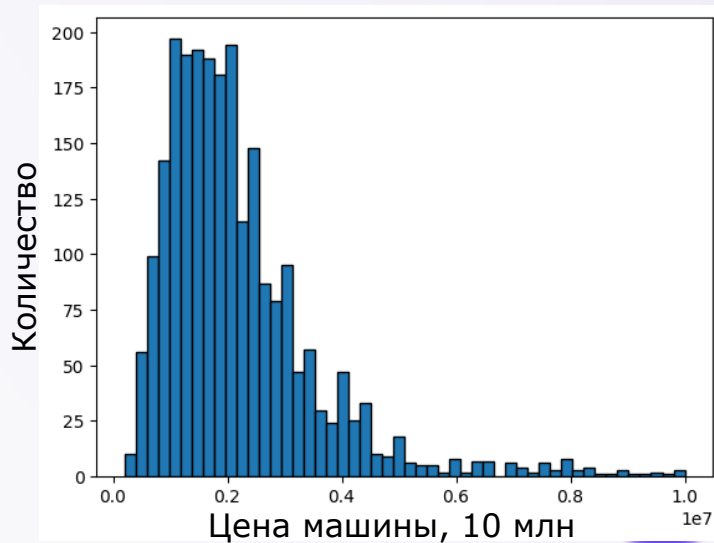


Убираем людей с машинами дороже 10 млн,
потому что они выбиваются из общего тренда

Корреляция численных параметров



По корреляционной таблице видно, что корреляция на хорошем уровне только у очевидной зависимости стажа вождения от возраста, поэтому мы стали искать неочевидные зависимости, такие как зависимость от стоимости машины



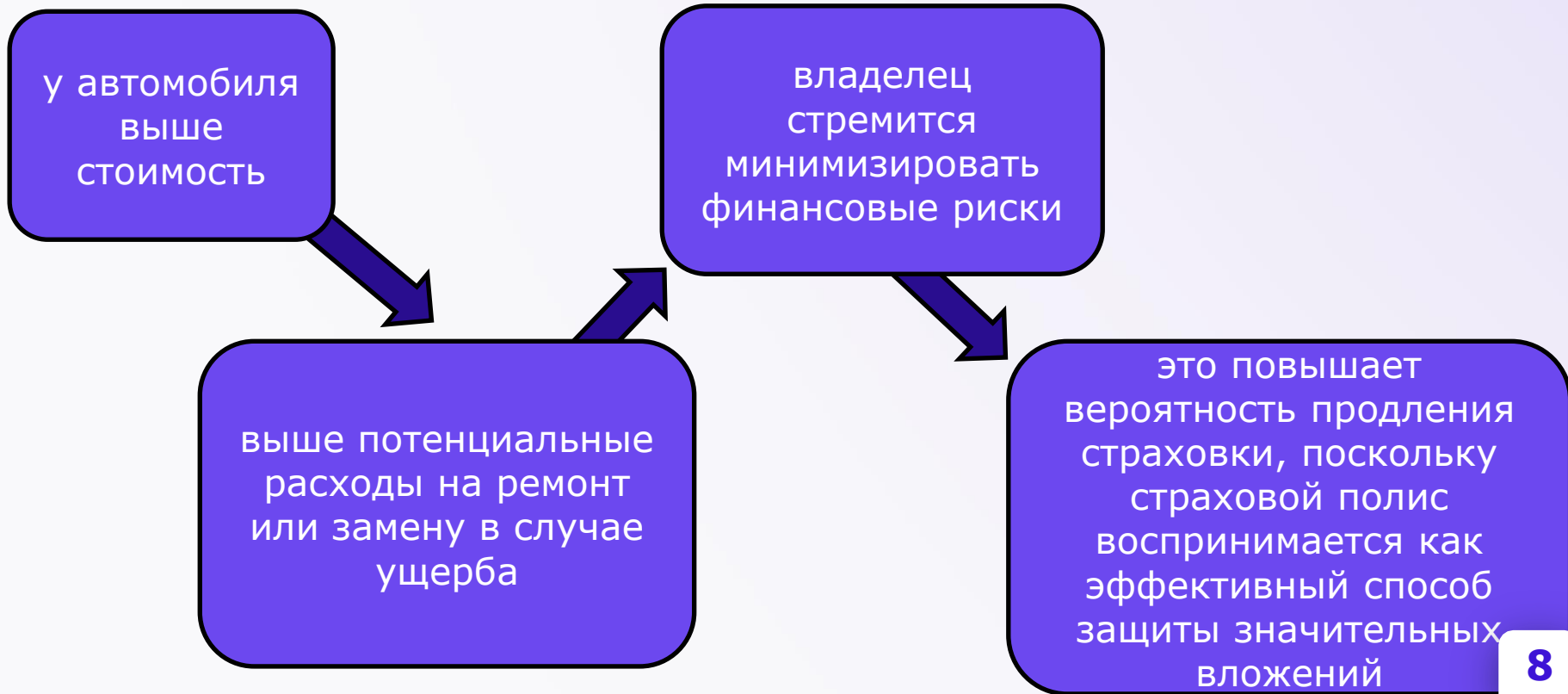
Исследовательский вопрос

Что может влиять на
желание клиента
застраховать авто?

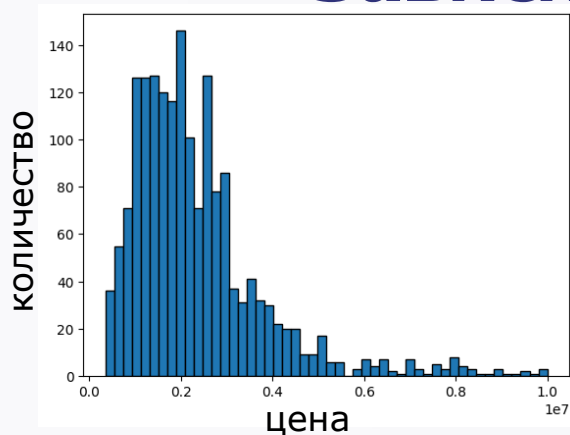
Гипотеза

Гипотеза состоит в том, что, чем выше цена автомобиля, тем больше шанс того, что автовладелец купит полис КАСКО

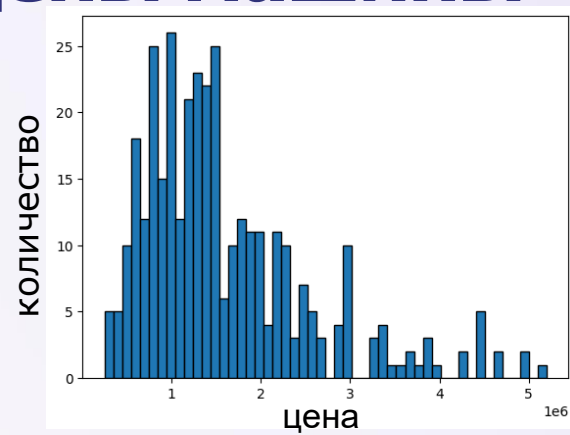
Механизм



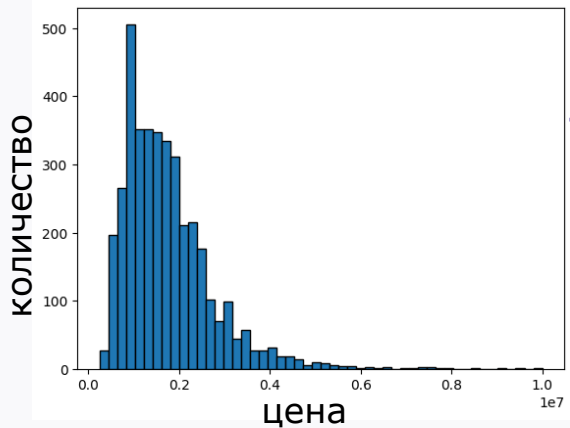
Зависимость от цены машины



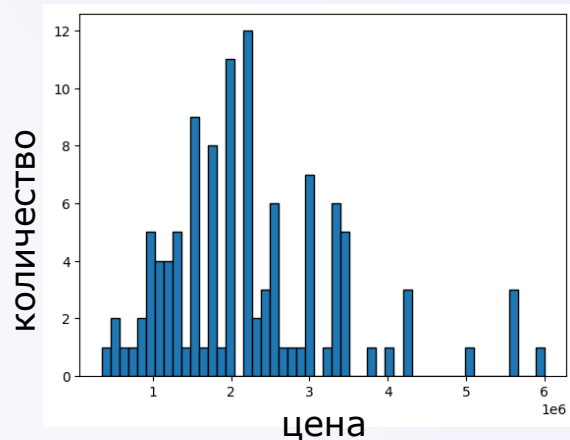
Полное КАСКО



Только угон



Тоталь + угон



Только ущерб

Зависимость от цены машины

t-тест Стьюдента показал p-value равное 0.365, а также коэффициент детерминации линейной регрессии равен 0.0004.

Это указывает на то, что зависимость шанса продления относительно стоимости машины минимальна

Итоги исследования

Гипотеза не подтвердилась

Ограничения: были предоставлены данные только одной компании в пределах одной страны

Применение исследования

Не стоит направлять рекламную кампанию на владельцев дорогих автомобилей, стоит быть одинаково ориентированными на владельцев автомобилей всех ценовых категорий.