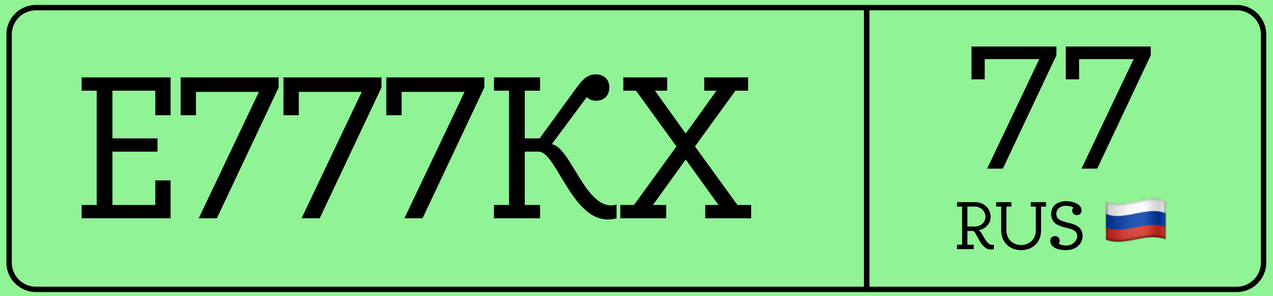
Н. НОВГОРОД



ЕДУ КАК ХОЧУ

DATA KOALAS

Информация о датасете

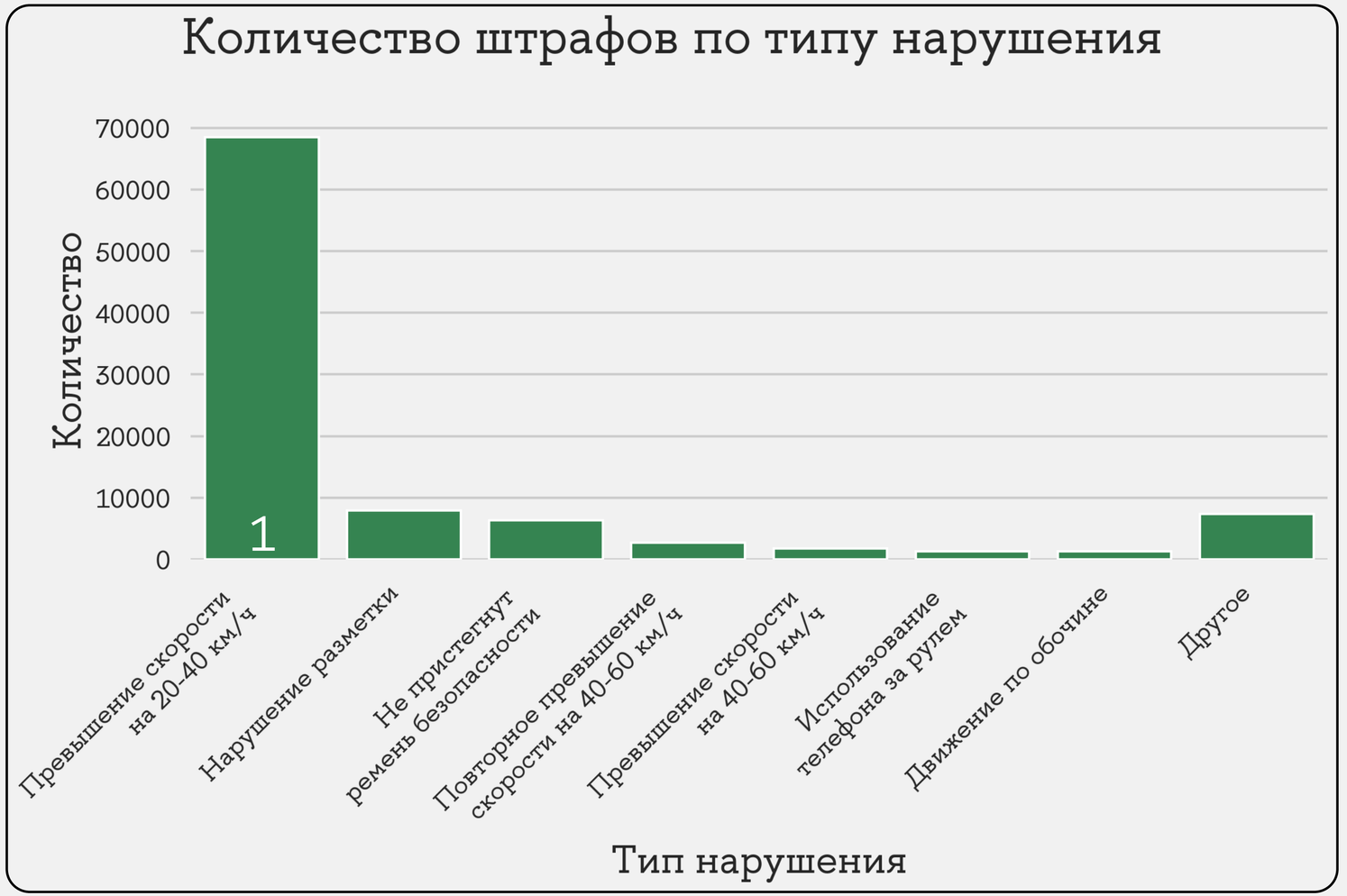
Статистика штрафов ГИБДД

- Данные по полученным клиентами Т-Банка штрафам ГИБДД
- **97 307** строк, **23** столбца
- Временной промежуток: с 28 апреля по 27 мая 2024 года

Предварительный анализ

Первым делом мы изучили, какие штрафы чаще всего получают.

Мода:
Превышение скорости на 20–40 км/ч



1. Нарушение скоростного режима является ключевым фактором риска смертности и травматизма на дорогах

Предварительный анализ

20-40 км/ч
несерьезное нарушение

40+ км/ч
серьезное нарушение

Формирование таргета

+ Создание переменной *target*

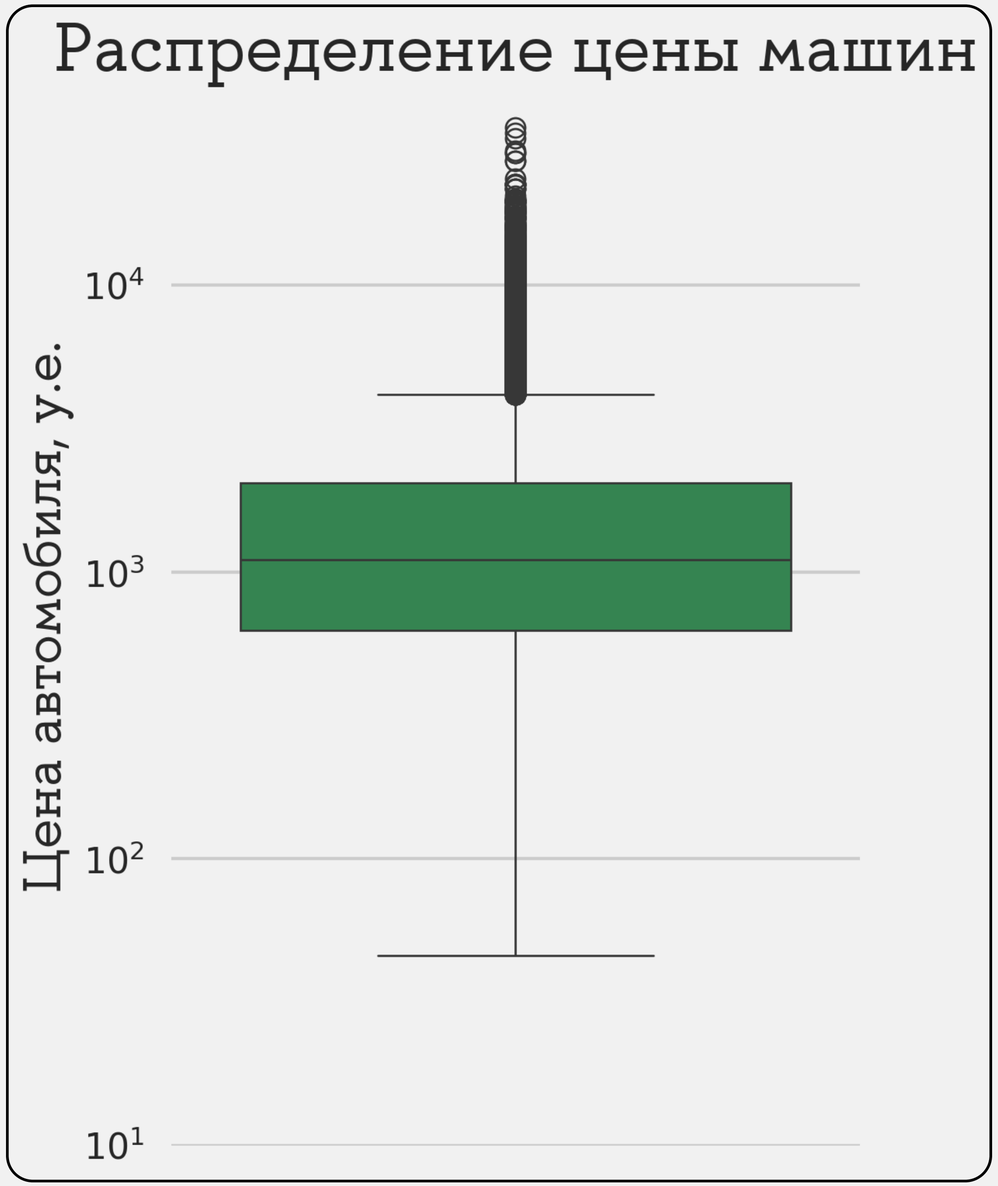
- 0, если превышение скорости на 20–40 км/ч
- 1, если превышение скорости больше, чем на 40 км/ч

Предварительный анализ

Среднее:
1615.32

Медиана:
1105.00

Стандартное отклонение:
1674.05



Очистка выбросов

- Удаляем некорректные значения по столбцам:
возраст, пол, год автомобиля, объем двигателя, цена машины, доход автомобилиста
- Удаляем NaN по цене машины

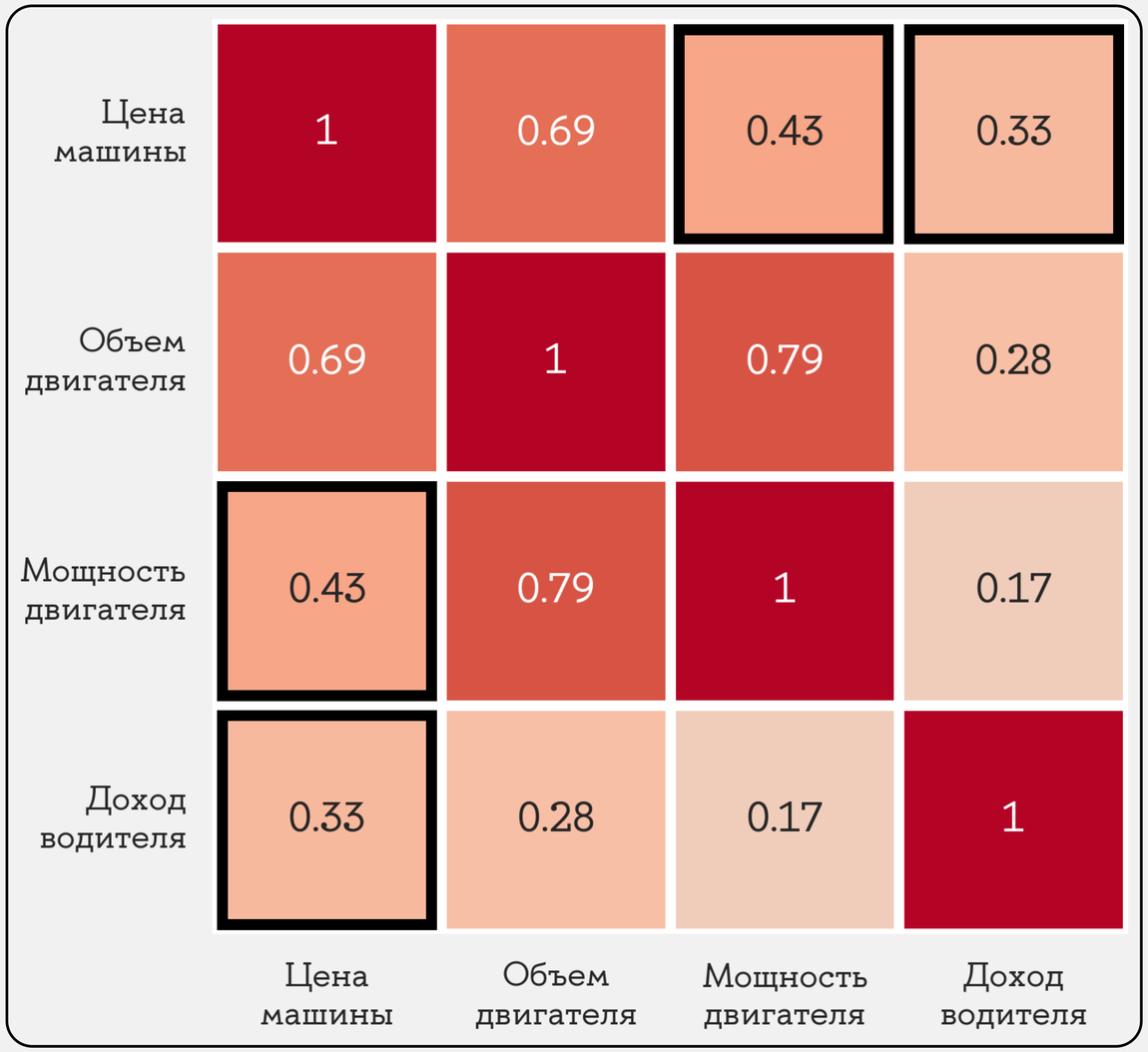
БЫЛО 97 307 → СТАЛО 84 070

- 13 237

- Оставляем по типам нарушения только превышение скоростного режима.
- Оставляем по типам автомобиля только легковые

БЫЛО 84 070 → СТАЛО 63 440

-20 630



Матрица корреляций

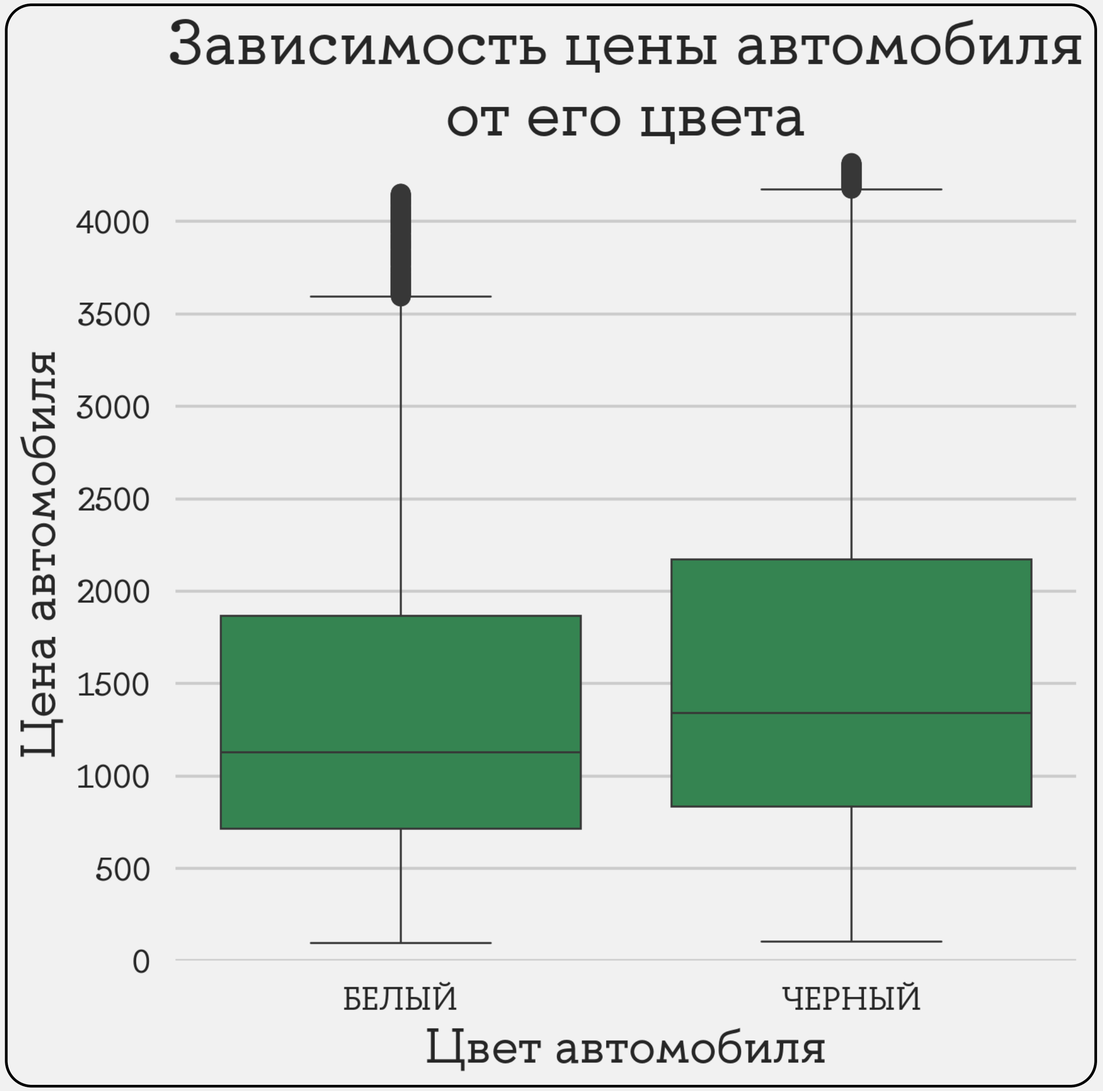
Полученные значения дают предпосылки для построения механизма и его проверки.

Поиск гипотезы

Мы решили поискать “забавные” зависимости в данных.

Т-критерий Уэлча
p-value = 0.002

Было выявлено, что черные автомобили в среднем стоят дороже.

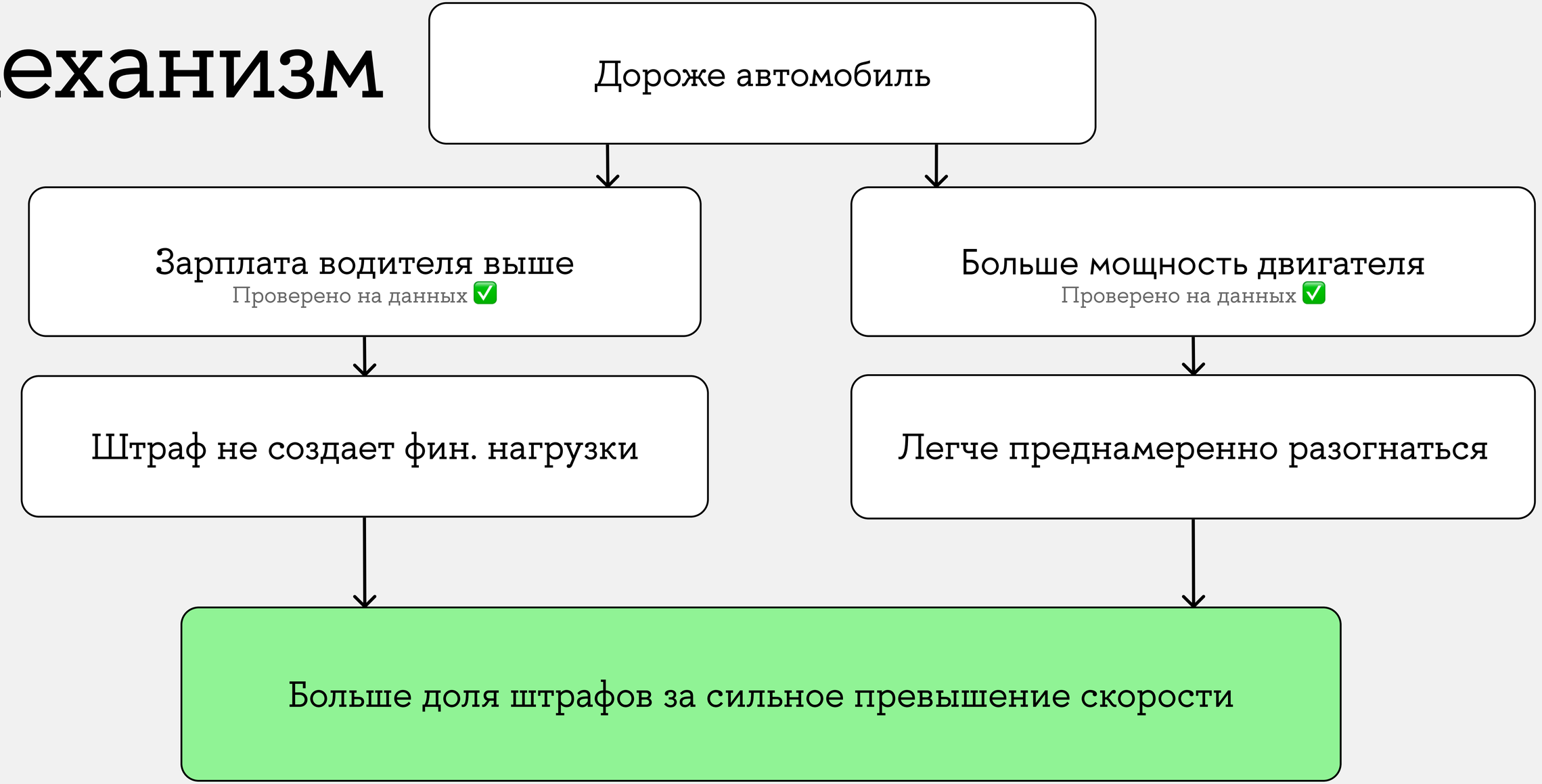


Исследовательский вопрос и гипотеза

Как связана цена автомобилей клиентов с характеристиками правонарушений?

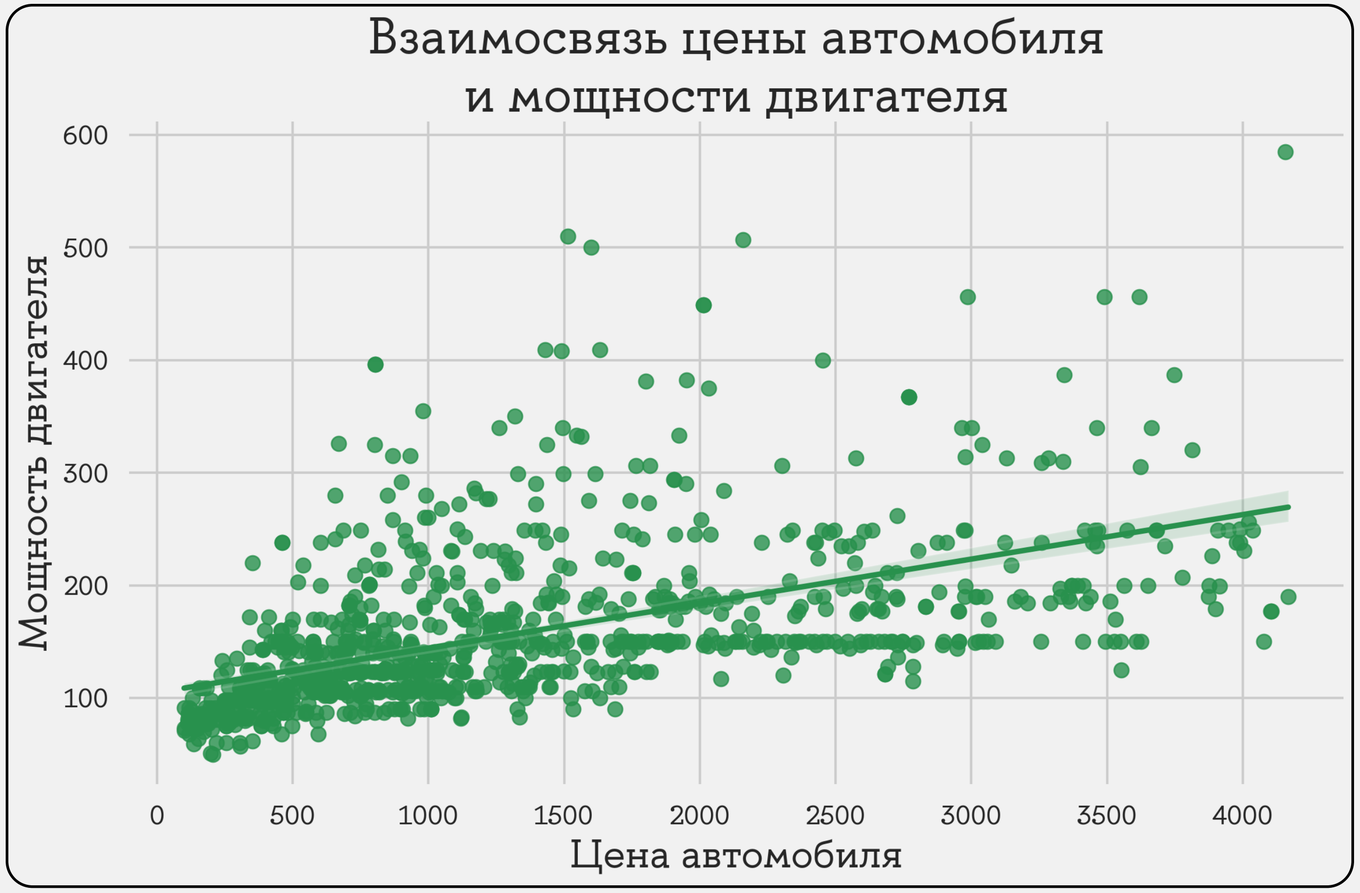
Чем дороже легковой автомобиль, тем выше доля серьезных превышений скорости среди всех случаев превышения.

Механизм


Дороже автомобиль
Зарплата водителя выше
Проверено на данных ✅
Больше мощность двигателя
Проверено на данных ✅
Штраф не создает фин. нагрузки
Легче преднамеренно разогнаться
Больше доля штрафов за сильное превышение скорости

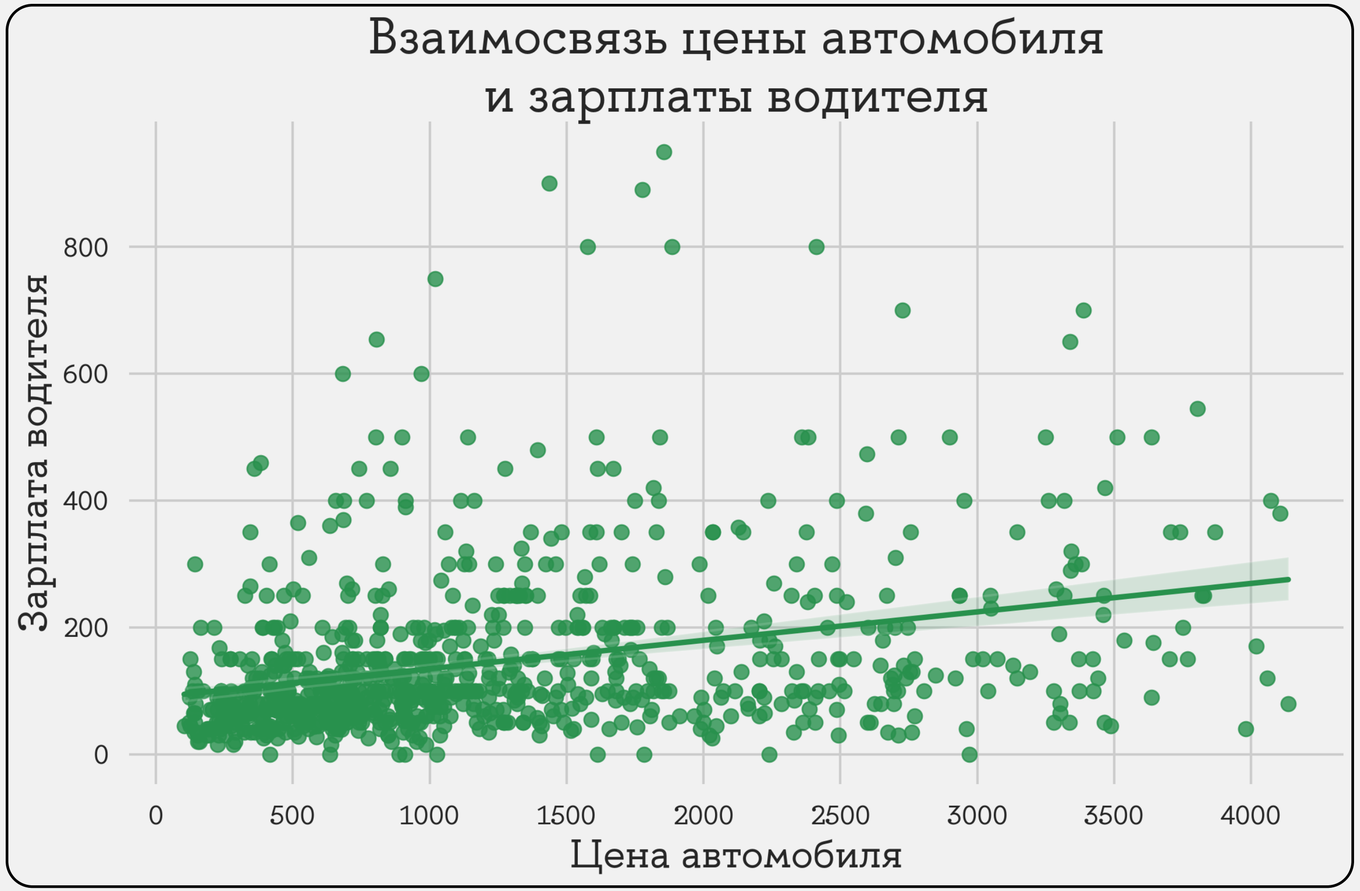
Проверка механизма

Линейная регрессия
engine_power от car_price:
p-value < 0.01
b_0: 103.58
b_1: 0.04
R^2 = 0.29



Проверка механизма

Линейная регрессия
engine_power от car_price:
p-value < 0.01
b_0: 94.82
b_1: 0.04
R^2 = 0.14



Математическая модель

Итоговый размер выборки:
64 440

Уровень статистической значимости:
0.05

Т-критерий и логистическая регрессия выбраны из-за того, что они хорошо подходят для бинарных данных и дополняют друг друга

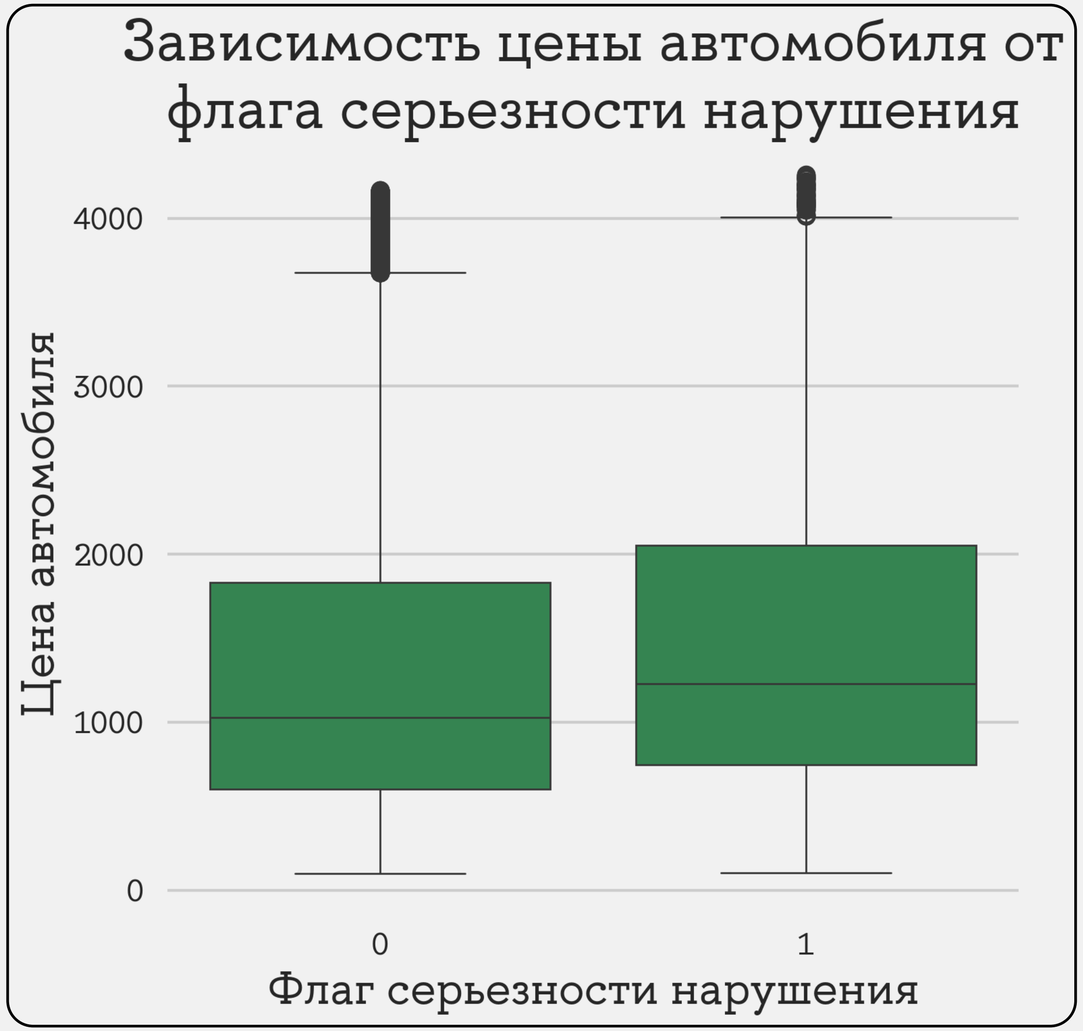
Для Т-критерия мы проверили дисперсии тестом Левене:
p-value < 0.01

Проверка моделей

T-test Уэлча:
p-value < 0.01

$$target_i \sim b_0 + b_1 \cdot carprice_i + \epsilon_i$$

Логистическая регрессия:
p-value < 0.01
b_0: -2.7327
b_1: 0.0008



Проверка устойчивости

Проверка по группам и доп. переменным

Разбиваем по дням недели и уровню образования

В каждой группе мы:

1. Строим логистическую регрессию и проводим T-test
2. Анализируем результат

Добавляем контрольные переменные

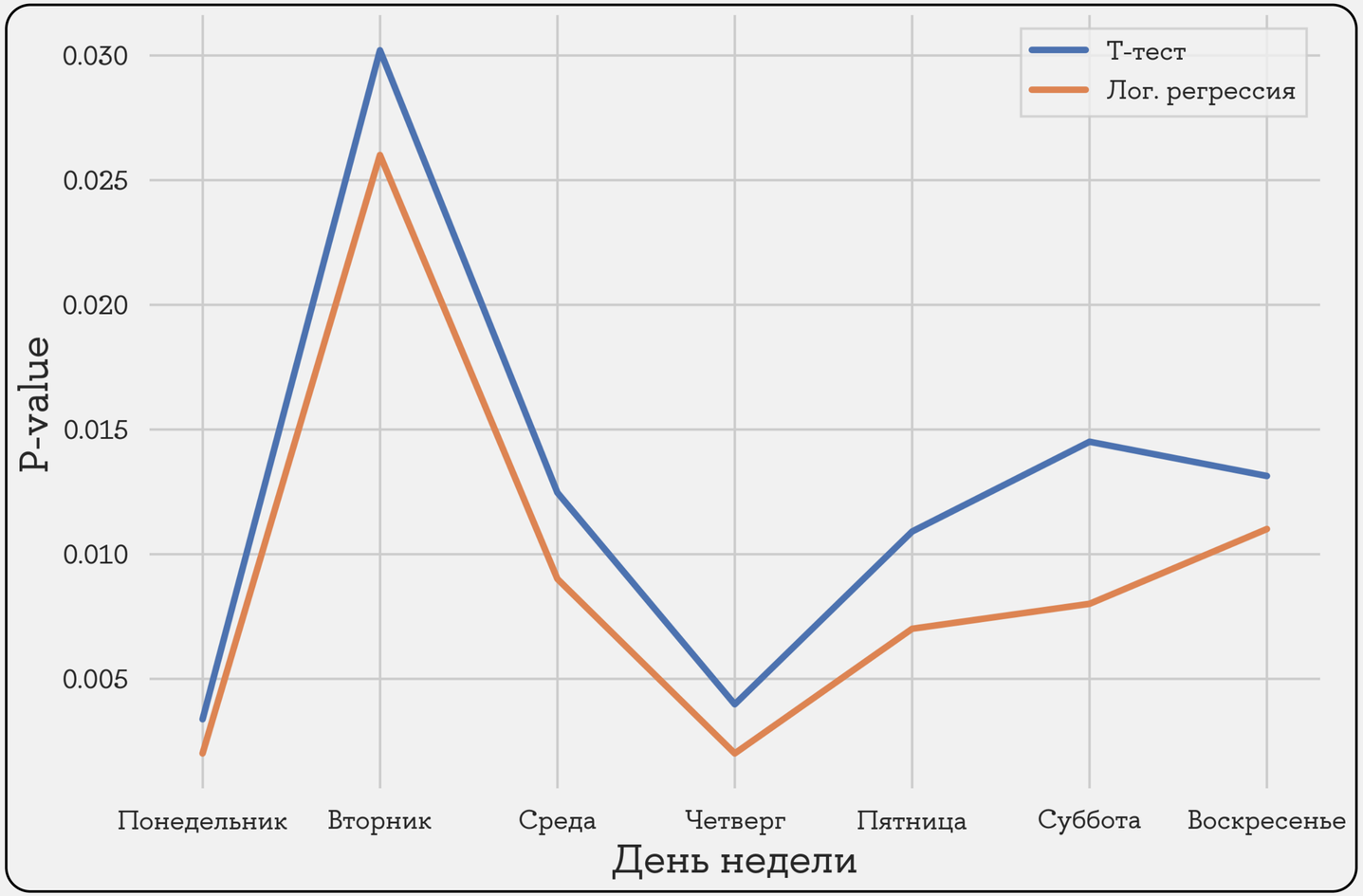
В каждой группе мы:

1. Строим множественную логистическую регрессию
2. Анализируем результат

Разбиваем по дням

День недели	Кол-во поз. примеров	Кол-во нег. примеров	Подтвердилась ли гипотеза
Понедельник	896	11412	✅
Вторник	546	8119	✅
Среда	545	8072	✅
Четверг	622	9187	✅
Пятница	765	10471	✅
Суббота	423	5313	✅
Воскресенье	511	6558	✅

График p-value по дням недели



Разбиваем по уровню образования

Уровень образования	Кол-во поз. примеров	Кол-во нег. примеров	Подтвердилась ли гипотеза
Высшее	1667	22315	✅
Начальное, среднее	885	11008	✅
Неполное высшее	501	6537	✅
Второе высшее	59	1084	✅
Ученая степень	18	243	❌

Слишком мало данных + Осознаность людей с учёной степенью выше

Добавляем контрольные переменные

Добавленные признаки	Год автомобиля, возраст, доход, пол, праздничный ли день
Результат	p-value < 0.01 Коэффициент при целевой переменой = 0.0002

Интерпретация

Гипотеза подтвердилась ✅

Чем дороже легковой автомобиль, тем выше доля серьезных превышений скорости среди всех случаев превышения.

Проведена проверка на устойчивость ✅

- На всех группах, кроме учёной степени, гипотеза подтвердилась.
- Интерпретировали почему гипотеза не подтверждается на группе людей с учёной степенью

Ограничения и перспективы

Маленький временной промежуток:
конец апреля — конец мая

→ Расширение временного интервала

Отсутствие данных по всем штрафам одного водителя ⇒ невозможно проанализировать динамику

→ Полная история штрафов для каждого водителя

Исследование распространяется только на клиентов Т-Банка

Альтернативные направления влияния на гипотезу

Расширенные характеристики региона:

- Плотность населения
- Другие социально-демографические признаки

Расширенные характеристики штрафа:

- Место совершения нарушения
- Нарушал ли до этого водетель
- Попадал ли водитель в аварии

Введение

Система day-fines в Финляндии

$$\text{Размер штрафа, €} = \frac{\text{Месячный доход} - €255}{60} \times \left(\begin{array}{c} \text{превышение сверх 23 км/ч} \\ \text{округленное вниз до четного} \end{array} + \begin{array}{l} 12 \text{ в городе} \\ 10 \text{ за городом} \end{array} \right)$$

Не применимо в России:
Многие (до трети населения) получают серую или черную заработную плату[1].

1. Ведомости, по информации от министра финансов, 2017

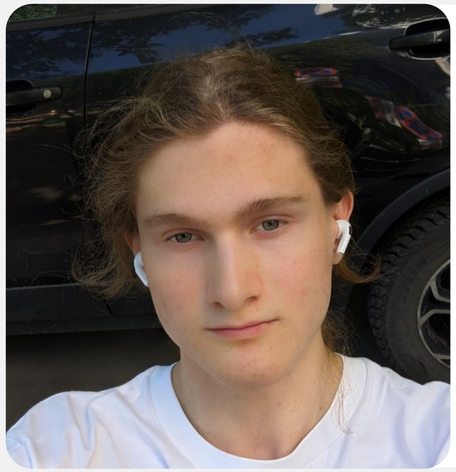
Policy implication

Государству	Т-Банк
Динамические штрафы Установить размер штрафа как долю от цены автомобиля. В качестве поставщиков данных о ценах автомобилей использовать страховые компании, которые учитывают различные характеристики (мощность двигателя, модель и год выпуска, пробег и так далее)	**Т-Страхование** Советовать страхование Каско пользователям, которые сильно превышают скоростной режим

наша команда

Данис Динмухаметов

11 КЛАСС
@SEYOLAX



Андрей Михайлов

11 КЛАСС
@LEGREY4IK

Герман Иванов

11 КЛАСС
@GERMANIVANOV0719

Арсений Чугунов

11 КЛАСС
@LOTOSSOKS

Андрей Хлопотных

11 КЛАСС
@ANDREYKHLOPOTNUKH

Приложение

приложение


ICELAND
FINLAND
NORWAY
SWEDEN
RUSSIA
DENMARK
UNITED KINGDOM
IRELAND
BELARUS
POLAND
BELGIUM
UKRAINE
AUSTRIA
FRANCE
ROMANIA
KAZAKHSTAN
MONGOLIA
ITALY
BULGARIA
KYRGYZSTAN
SPAIN
GREECE
TURKEY
TURKMENISTAN
NORTH KOREA
SOUTH KOREA
JAPAN
CHINA
TUNISIA
SYRIA
IRAQ
IRAN
AFGHANISTAN
MOROCCO
JORDAN
PAKISTAN

Проверка полиномиальности

Логарифм

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.6687	0.301	-8.860	0.000	-3.259	-2.078
car_price	0.0001	4.25e-05	2.749	0.006	3.35e-05	0.000
car_price_ln	-0.0113	0.051	-0.223	0.824	-0.111	0.088

Квадрат

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7704	0.046	-60.233	0.000	-2.861	-2.680
car_price	0.0002	6.2e-05	2.686	0.007	4.5e-05	0.000
car_price^2	-1.602e-08	1.63e-08	-0.980	0.327	-4.81e-08	1.6e-08

Корень

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.7366	0.030	-90.589	0.000	-2.796	-2.677
car_price	0.0001	1.84e-05	5.815	0.000	7.11e-05	0.000
car_price^0.5	6.728e-05	0.001	0.132	0.895	-0.001	0.001

Проверка выборочных средних car_price для T-критерия

p-value Shapiro = 0.51