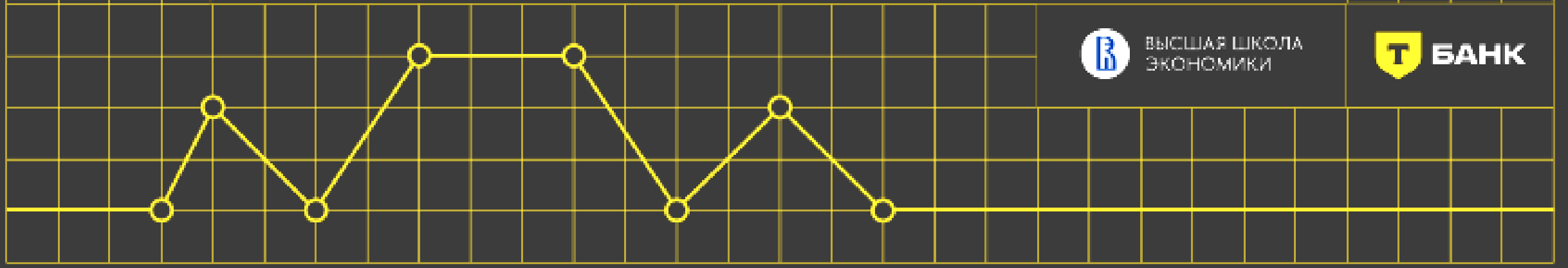
DANO

ИССЛЕДОВАНИЕ

©18 команда

DANO HAKATON & HSE NN



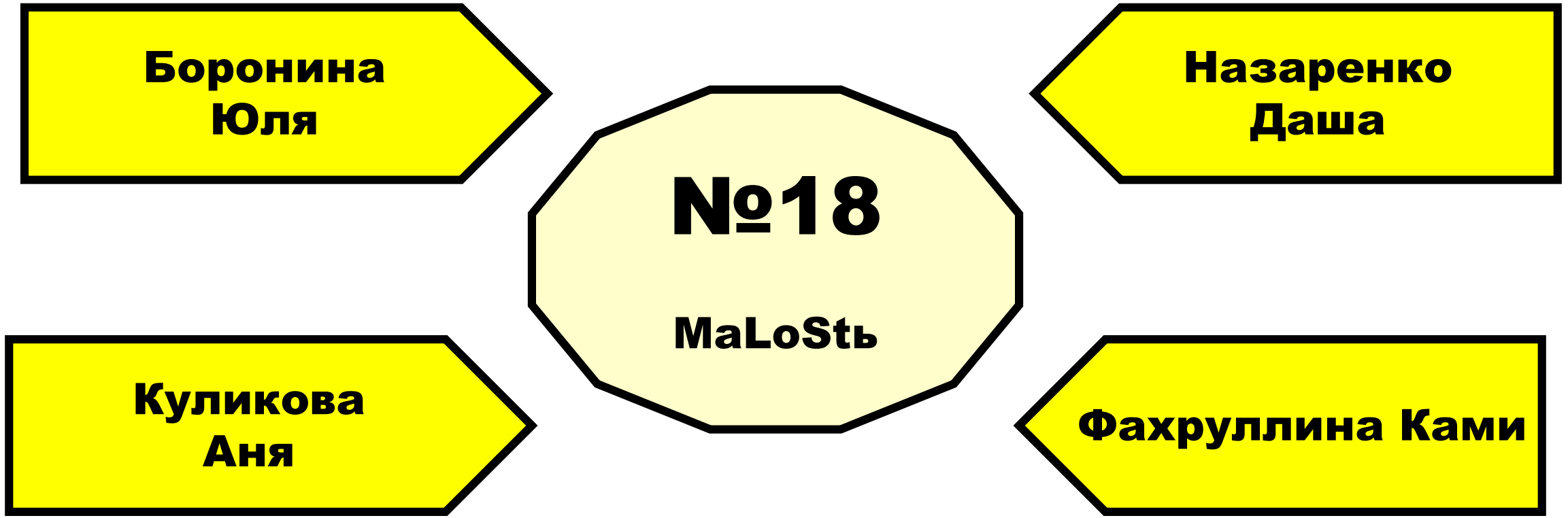
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

БАНК

01

Предварительный анализ данных

Data Scientисты



Math+Logic+Statistics

Структура базы данных

! часть данных (~0,5% от полученных в каждый день) по полученным клиентами Т-Банка штрафам ГИБДД* - 97 308 строк, 22 характеристики

Период: с 28 апреля 2024 г по 28 мая 2024 г.

---

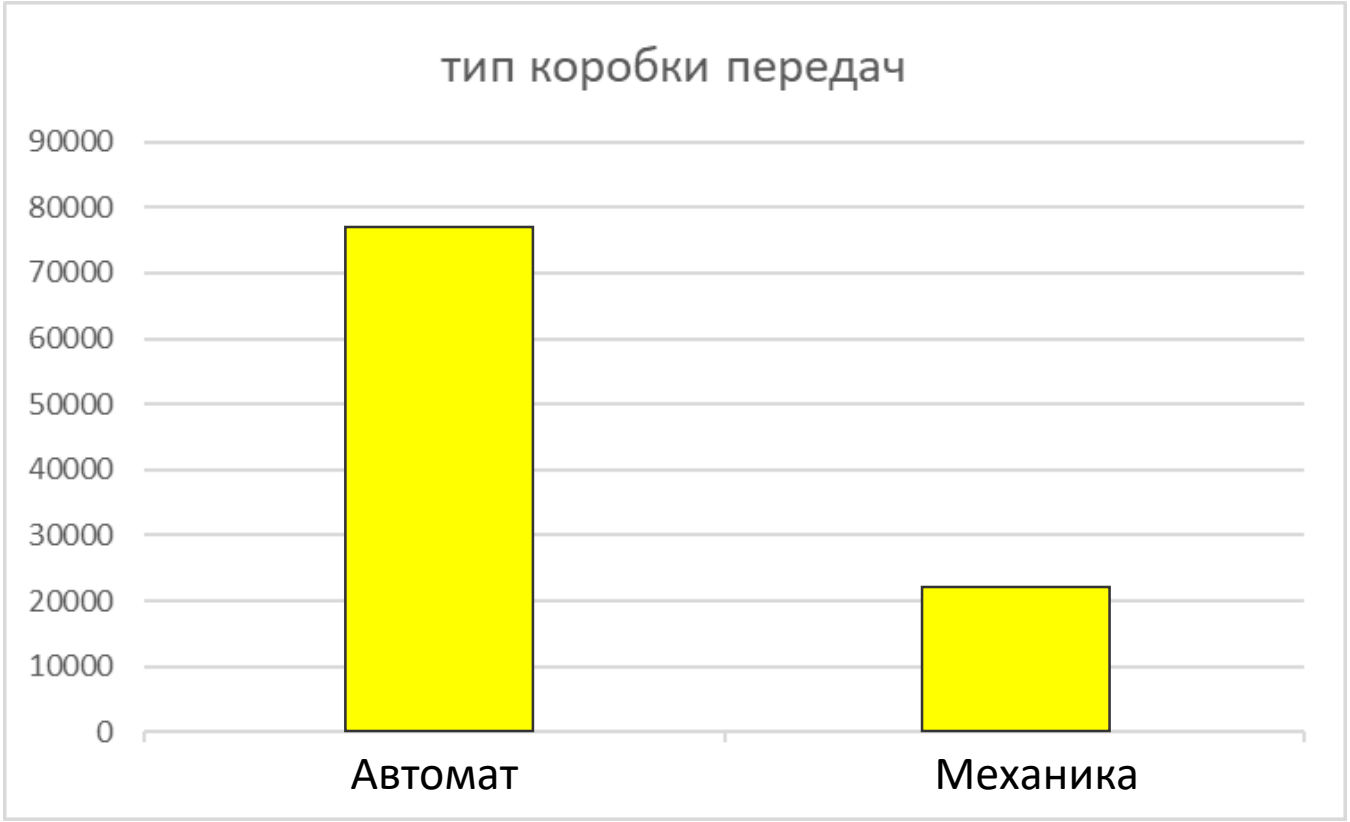
*Одна запись – один уникальный штраф ГИБДД. К характеристикам штрафа добавлена информация о характеристиках водителя и автомобиля на котором совершено правонарушение.

Структура базы данных

Структура базы данных

+34 типа малозначимы

Структура базы данных



Структура базы данных

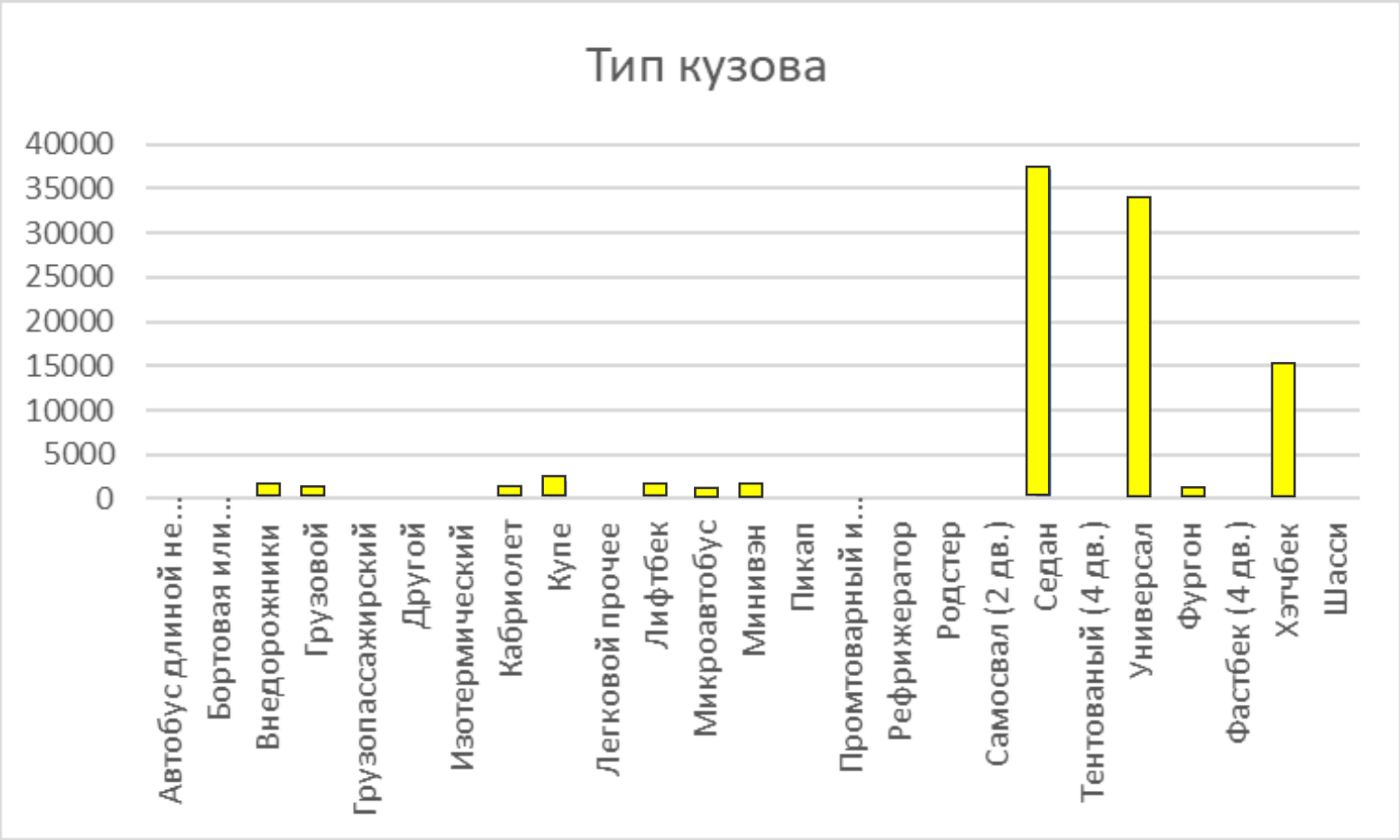
Структура базы данных

Структура базы данных

Уникальные водители: 1297 (с превышением скорости более 5 раз)

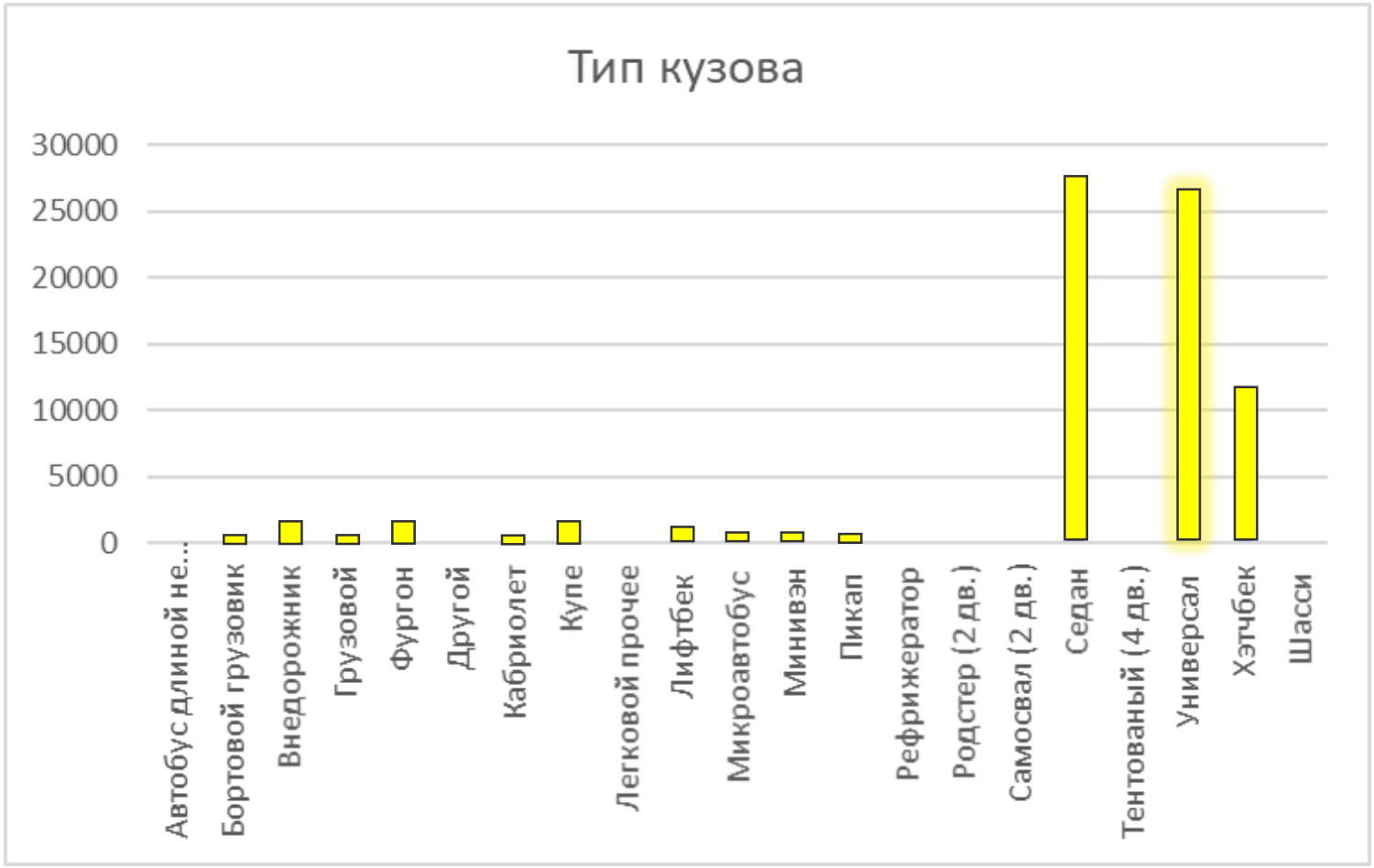
Структура базы данных

Количество штрафов с повторениями



Структура базы данных

Количество штрафов без повторений (=количество кузовов)



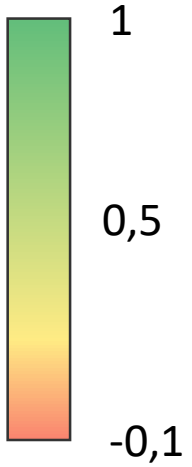
БЫЛО: 9728

СТАЛО: 73555

Структура базы данных

Предварительный анализ с помощью корреляционной таблицы

	engine_power	*car_price*	*age*	*children_cnt*	*income*
engine_power	1	0,546196121	0,04852	0,004078043	0,166865959
car_price	0,546196121	1	0,10217	0,005724679	0,181820769
age	0,048522478	0,102166076	1	0,118170086	-0,015317977
children_cnt	0,004078043	0,005724679	0,11817	1	0,024494617
income	0,166865959	0,181820769	-0,01532	0,024494617	1



Зависимости количественных величин:

engine_power – мощность двигателя автомобиля
car_price – стоимость автомобиля (условные единицы)
age – возраст владельца
children_cnt – количество детей у владельца
income - размер месячного дохода водителя

02

Постановка гипотезы №1

Исследовательский вопрос		Гипотеза
Взаимосвязаны ли характеристики клиентов и их автомобилей с характеристиками правонарушений (время/дата совершения правонарушения, статья правонарушения) и если да, то как?	→	Соотношение частотности различных категорий штрафов у взрослых с детьми и у взрослых без детей отличаются

Данные | Гипотеза | Обработка | Результаты | Заключение

МЕХАНИЗМ

Рождение ребенка значимым шагом для семьи, что заставляет чувствовать больше ответственности

Ответственные водители реже делают то, что может угрожать жизни: нарушают скорость, проезжают на красный и т.д.

Для таких водителей-родителей характерны нарушения, связанные с остановкой, заездом за линию и т.д.

03

Проверка гипотезы


Сравнение количества штрафов в зависимости от наличия детей
%
80
70
60
50
40
30
20
10
0
Нарушение разметки
Не пристегнут ремень...
Не пропустил пешехода
Остановка или стоянка в месте,...
Остановка или стоянка в...
Остановка или стоянка на...
Остановка или стоянка на...
Остановка или стоянка на...
Остановка на перекрестке
Пересечение стоп-линии
Поворот не из крайней полосы
Повторное превышение...
Повторный выезд на полосу...
Повторный проезд на...
Превышение
Ряд1
Ряд2

Сравнение количества штрафов в зависимости от наличия детей

t-тест

P-value = 0,09

Гипотеза **не** подтверждается

Гипотеза опровергнута

03

Результаты анализа

Гипотеза

Водители совершают схожие нарушения в одинаковом количестве.

Наличие детей (как психологический фактор) не влияет.

Практическая значимость

Службам ГИБДД	• Проведение психологических тренингов для родителей

Ограничения и перспективы

Ограничения	Перспективы
• Номера владельцев, а не непосредственно владельцы	• Получить данные о конкретных владельцах через систему • Протестировать социальные критерии на большей выборке • Провести эмпирическое исследование о вождении автомобилей с типом кузова «Универсал»

Каких данных не хватило?

- Стаж вождения
- Возраст ребенка (детей)
- Дата покупки машины

Количественных данных

Спасибо!
У нас всё

Тест Фишера

Статистическая значимость – это количественный показатель, указывающий на то, что полученные результаты не являются случайными и могут быть признаны достоверными.

Она часто используется в маркетинге, когда нужно **проверить чистоту эксперимента и понять, можем ли мы доверять результатам теста.**

02

Постановка гипотезы №2

Рекомендации

Производителям

- Внедрять передовые технологий безопасности (уведомления о превышении скорости, встроенный навигатор)
- Обучение водителей (привыкание к новой коробке передач)
- Работа над репутацией
- Сотрудничество с органами ГБДД
- Улучшение качества сборки

Исследовательский вопрос		Гипотеза
Взаимосвязаны ли характеристики клиентов и их автомобилей с характеристиками правонарушений (время/дата совершения правонарушения, статья правонарушения) и если да, то как?		Время выписки штрафа зависит от наличия детей в семье: у человека, который имеет детей, большее количество штрафов будет приходится на отличающееся время

Данные | Гипотеза | Обработка | Результаты | Заключение

МЕХАНИЗМ

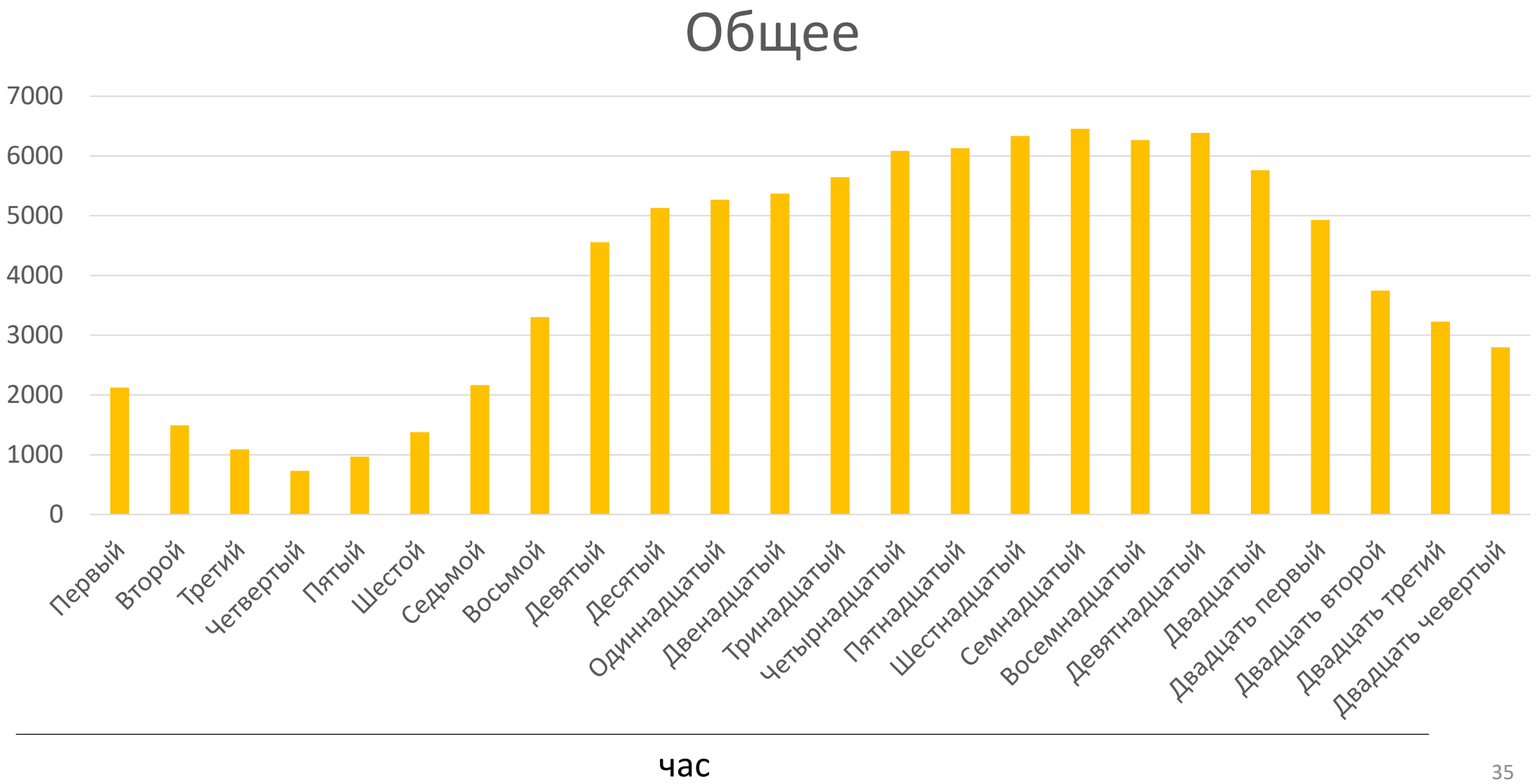
Дети посещают различные развивающие секции и кружки, ходят в школу и детский сад

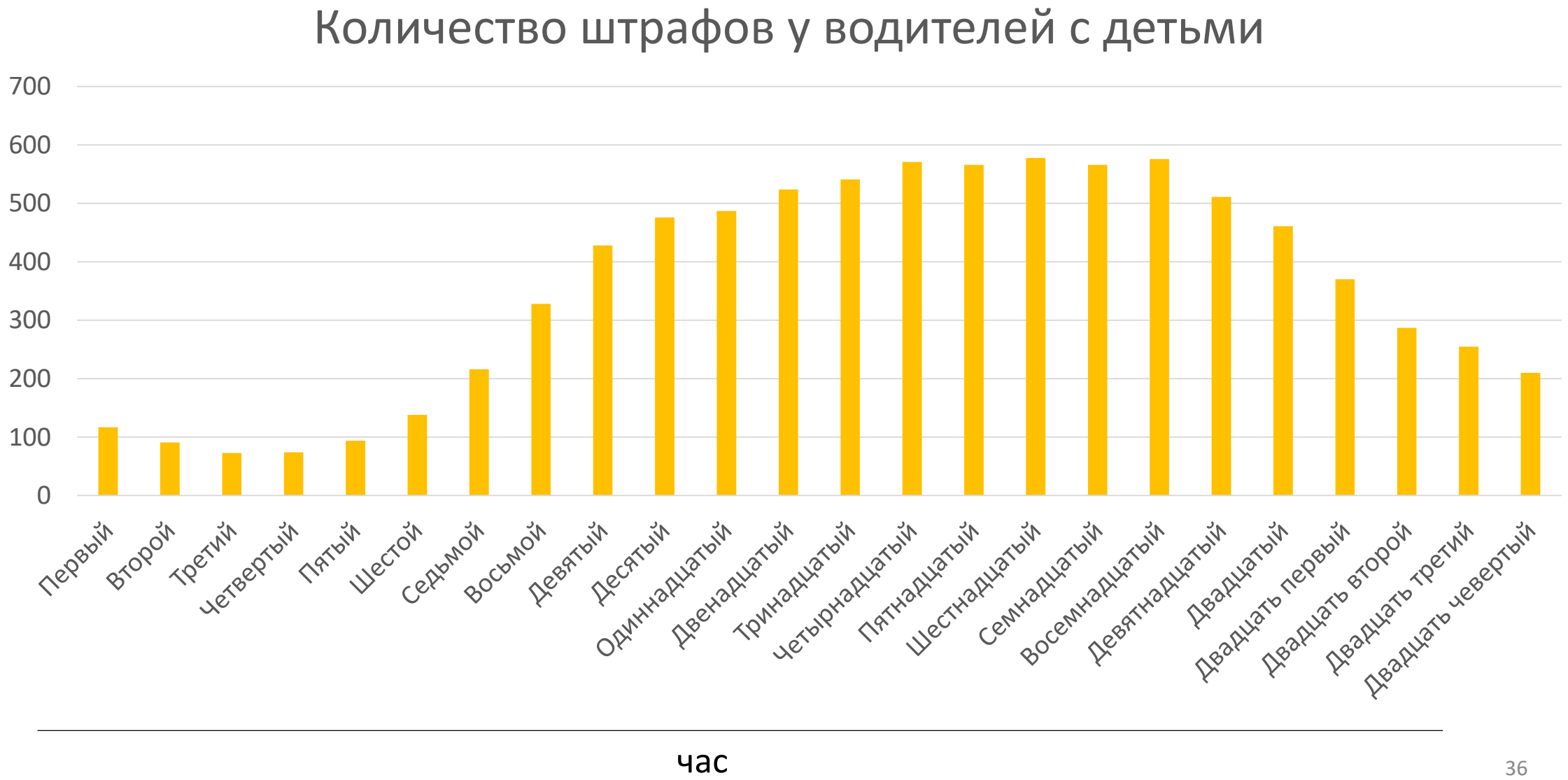
Многие места находятся вдали от дома, из-за чего родителям нужно возить ребенка, чтобы убедиться в его безопасности на пути

Поэтому возрастает необходимость возить ребенка на автомобиле в утренние и вечерние часы

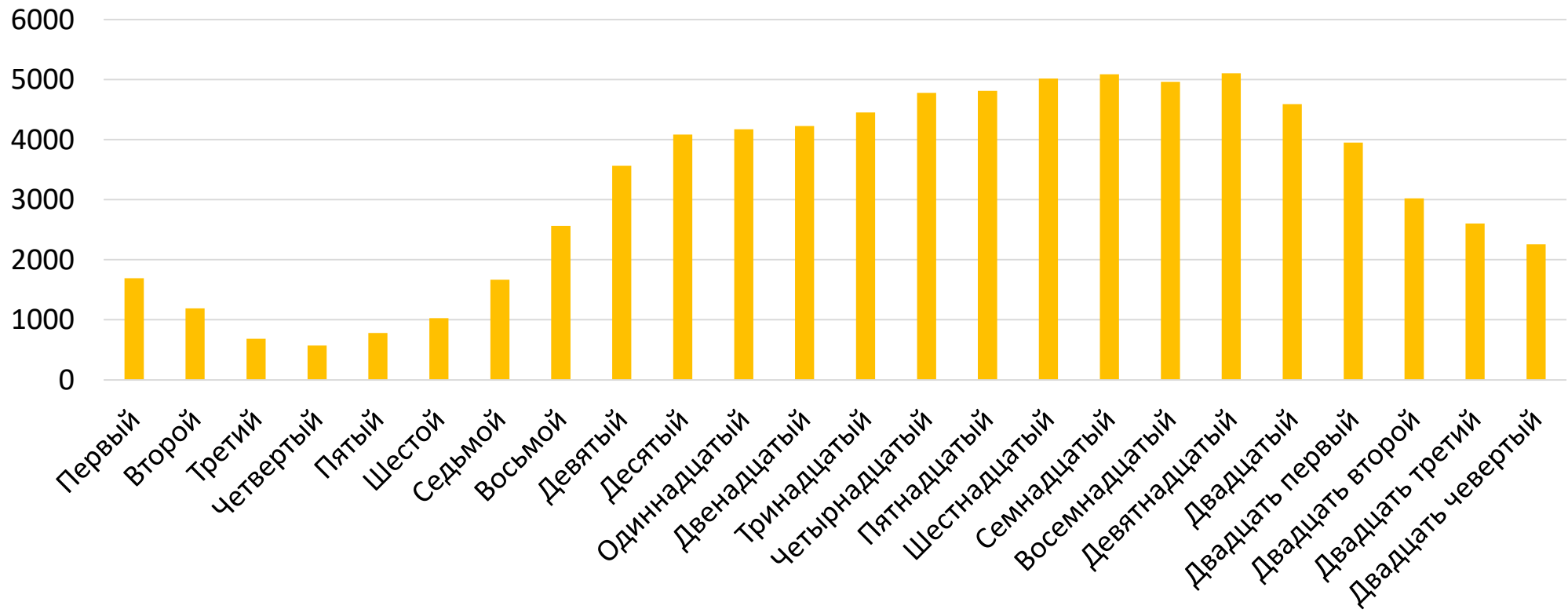
03

Проверка гипотезы №2


Общее
7000
6000
5000
4000
3000
2000
1000
0
Первый
Второй
Третий
четвертый
Пятый
Шестой
Седьмой
Восьмой
Девятый
Десятый
Одиннадцатый
Двенадцатый
Тринадцатый
четырнадцатый
Пятнадцатый
Шестнадцатый
Семнадцатый
Восемнадцатый
Девятнадцатый
Двадцатый
Двадцать первый
Двадцать второй
Двадцать третий
Двадцать чевертый
час


Количество штрафов у водителей с детьми
700
600
500
400
300
200
100
0
Первый
Второй
Третий
Четвертый
Пятый
Шестой
Седьмой
Восьмой
Девятый
Десятый
Одиннадцатый
Двенадцатый
Тринадцатый
Четырнадцатый
Пятнадцатый
Шестнадцатый
Семнадцатый
Восемнадцатый
Девятнадцатый
Двадцатый
Двадцать первый
Двадцать второй
Двадцать третий
Двадцать чевертый
час

Количество штрафов у водителей без детей



час

Коэффициент корреляции

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{S(x) \cdot S(y)} = \frac{1425847.708 - 355.75 \cdot 3202.5}{185.555 \cdot 1578.947} = 0.978$$

связь между весьма высокая и прямая

Маленький размер выборки:

Точный тест Фишера

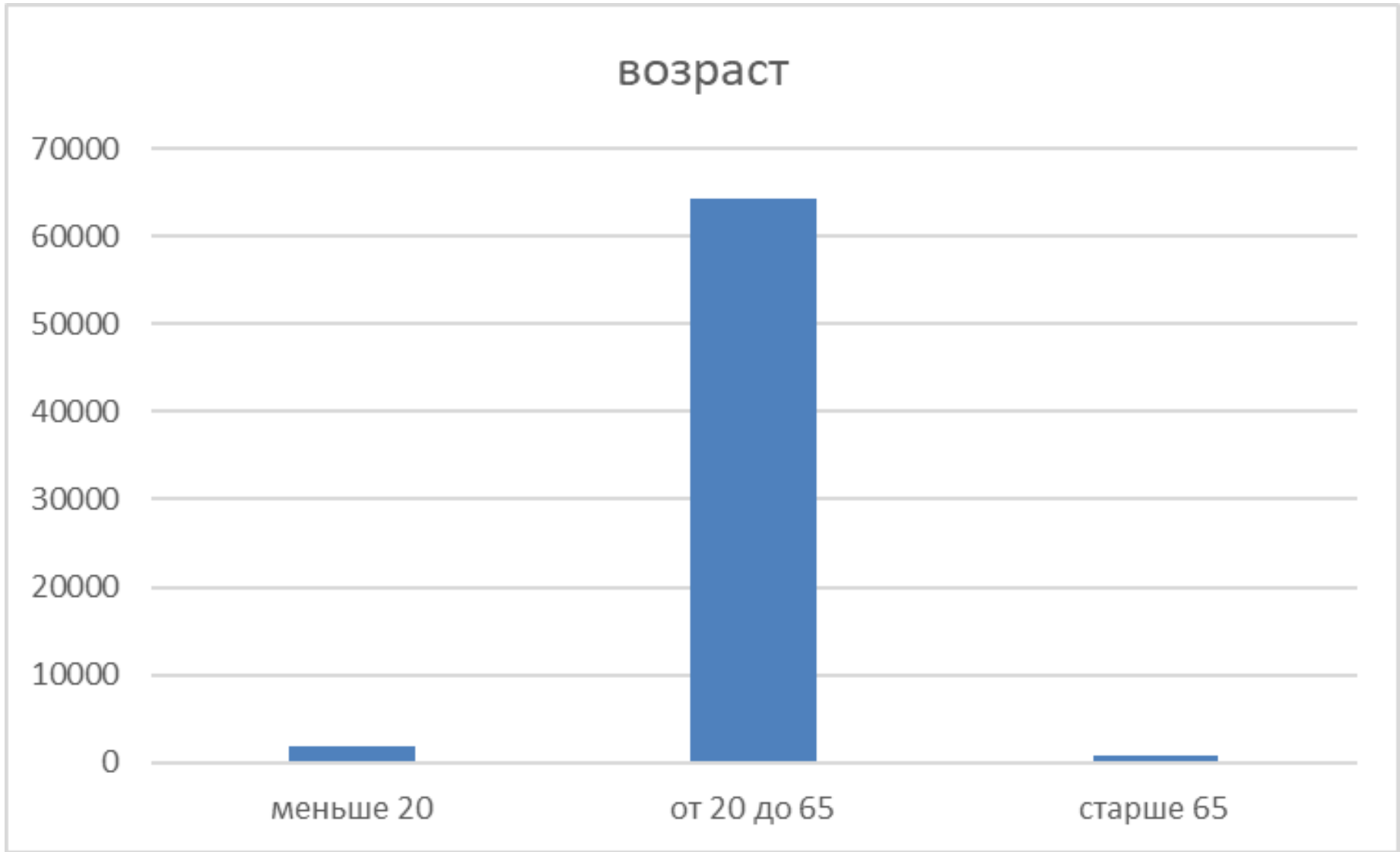
$$F = \frac{\sum (y_x - \overline{y})^2}{\sum (y_i - y_x)^2} \frac{n - m - 1}{m} = \frac{57238946.5551}{2594797.44} \cdot \frac{24 - 1 - 1}{1} = 485.301$$

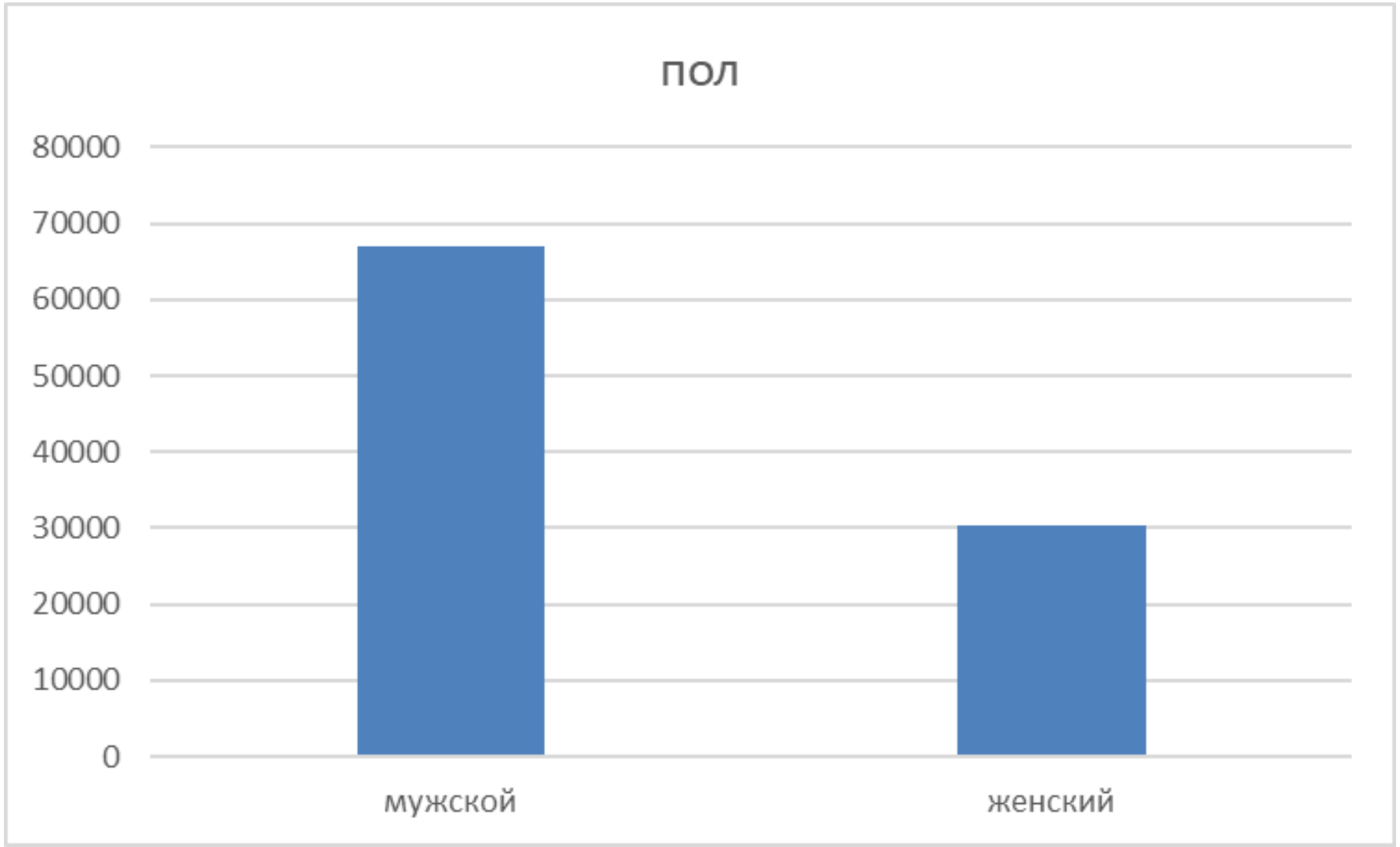
$$F_{табл} = 4.3009$$

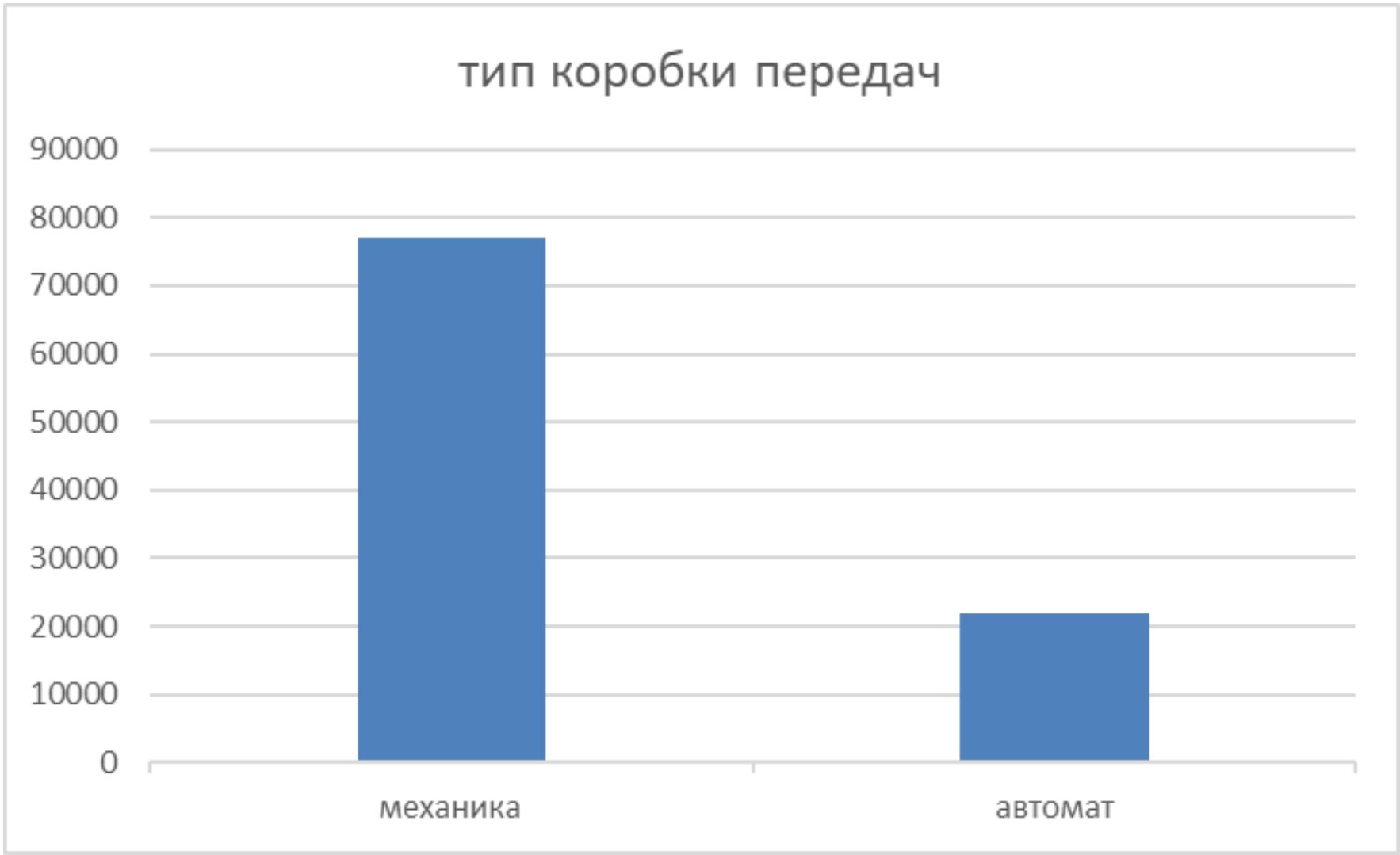
$$F > F_{табл}$$

найденная оценка уравнения регрессии статистически надежна

Гипотеза №2 опровергнута


возраст
70000
60000
50000
40000
30000
20000
10000
0
меньше 20
от 20 до 65
старше 65


пол
80000
70000
60000
50000
40000
30000
20000
10000
0
мужской
женский


тип коробки передач
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
механика
автомат

Регрессионный анализ

$$y_x = r_{xy} \cdot \frac{x - \overline{x}}{S(x)} \cdot S(y) + \overline{y} = 0.993 \frac{x - 3.571}{13.186} 12.496 + 3.571 = 0.941 \cdot x + 0.212$$

Связь прямая

Маленький размер выборки:

Точный тест Фишера

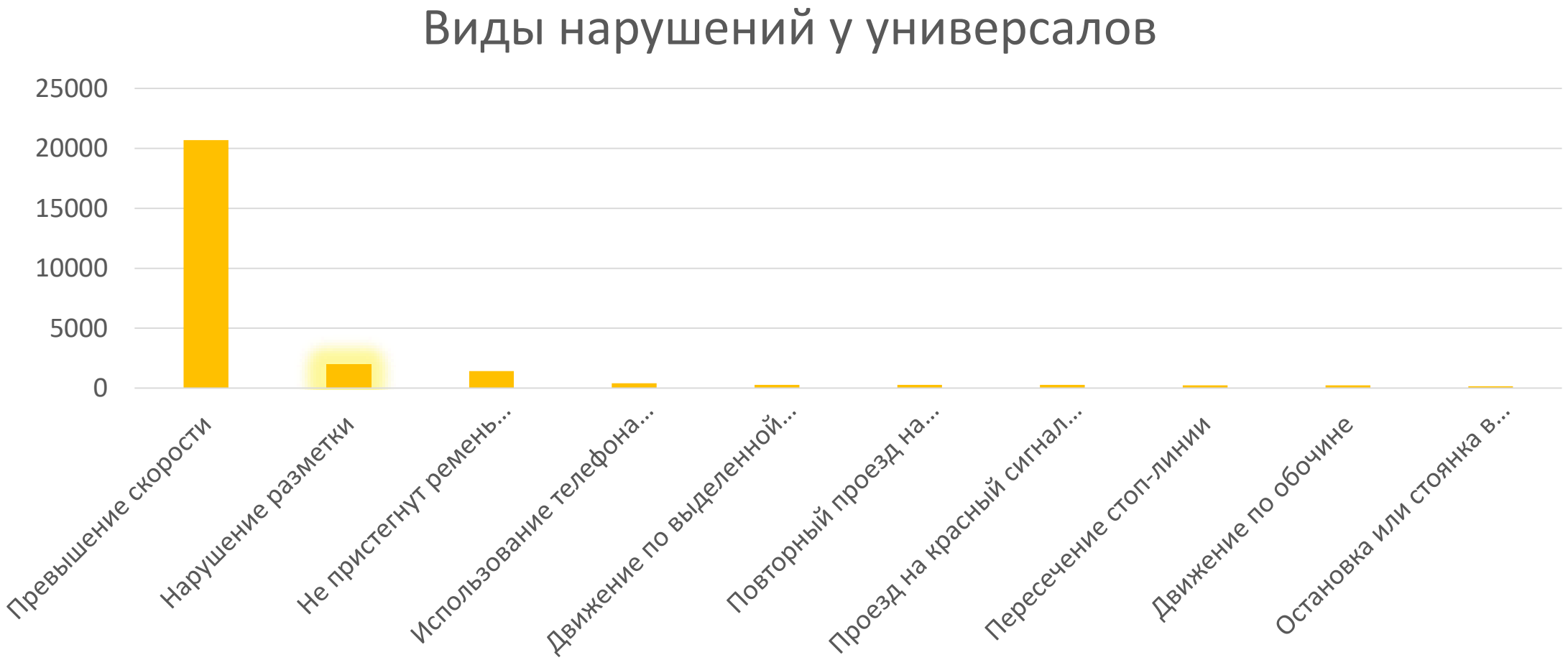
$F > F_{табл}$

коэффициент детерминации статистически значим (найденная оценка статистически надежна)

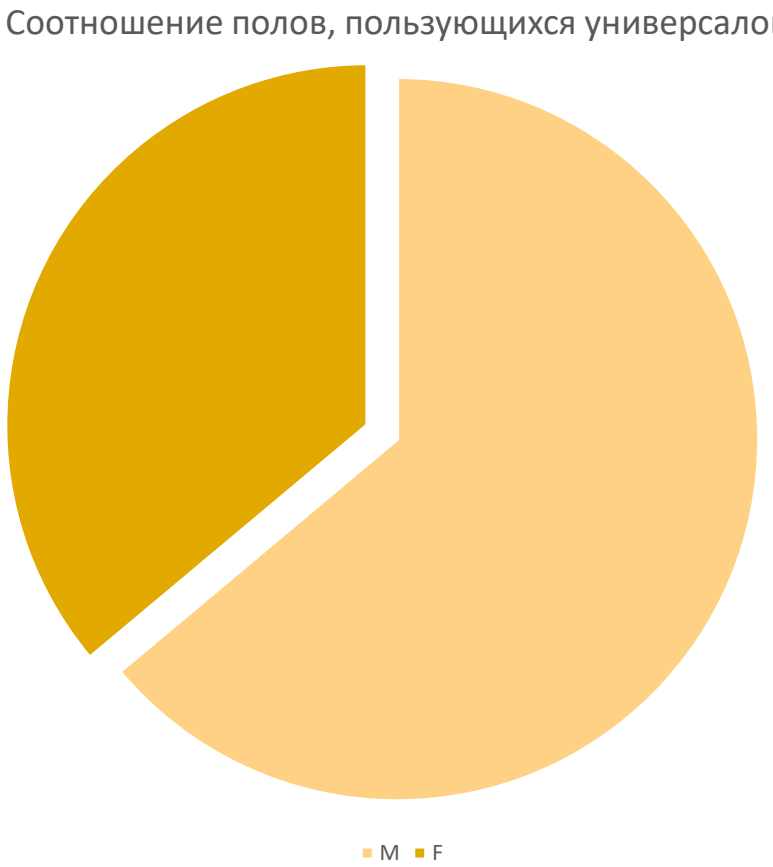
$$F = \frac{0.9852}{1 - 0.9852} \frac{28 - 1 - 1}{1} = 1729.152$$

$$F_{табл} = 4.2252$$

Структура базы данных



Структура базы данных



02

Постановка гипотезы №2

Исследовательский вопрос		Гипотеза
Взаимосвязаны ли характеристики клиентов и их автомобилей с характеристиками правонарушений (время/дата совершения правонарушения, статья правонарушения) и если да, то как?		Из самых распространенных кузовов автомобилей кузовы «Универсал» имеют больше нарушений

Большие кузовы имеют большую популярность за счет удобства перевозки вещей

При покупке автомобиля с кузово многие думают о комфорте перевозки, но не о неудобствах управления

Из-за частотными случаями аварий являются нарушения разметки, так как камера считывает малейшие заступы

МЕХАНИЗМ

02

Проверка гипотезы №3

Гипотеза №3 подтверждена