ИССЛЕДОВАНИЕ

Хахахахахакатон

1. Демидов Илья
2. Толстобров Артем
3. Екатерина Комкова
4. Озийгит Алекс
5. Нестеров Алексей
6. Худокормова Мария

Мотивация

Повлиять на часть общества, которая не перестает нарушать ПДД, несмотря на множество штрафов.

Исследовательский вопрос:

Взаимосвязаны ли характеристики клиентов и их автомобилей с характеристиками правонарушений и если да, то как?

Гипотеза:

Доля рецидивистов* среди богатых больше, чем среди бедных

*рецидивист – человек, нарушивший ПДД более **одного** раза

Механизм

Человек нарушает ПДД и платит штраф

Для бедных

штраф является **значительной** суммой денег относительно заработка

Человек старается водить аккуратнее и не нарушает правила

Для богатых

штраф является **незначительной** суммой денег относительно заработка

Человек никак не отреагировал

База данных

97307

штрафов, полученных
клиентами Т-Банка

В период с 28.04.2024 по 28.05.2024

Структура данных

Количественные

- Возраст водителя
- Размер месячного дохода водителя
- Объем двигателя автомобиля
- Мощность двигателя
- Цена машины
- Год выпуска автомобиля
- Время совершения правонарушения

Качественные

- Регион, в котором совершено правонарушение
- Пол водителя

Доработка датасета

923

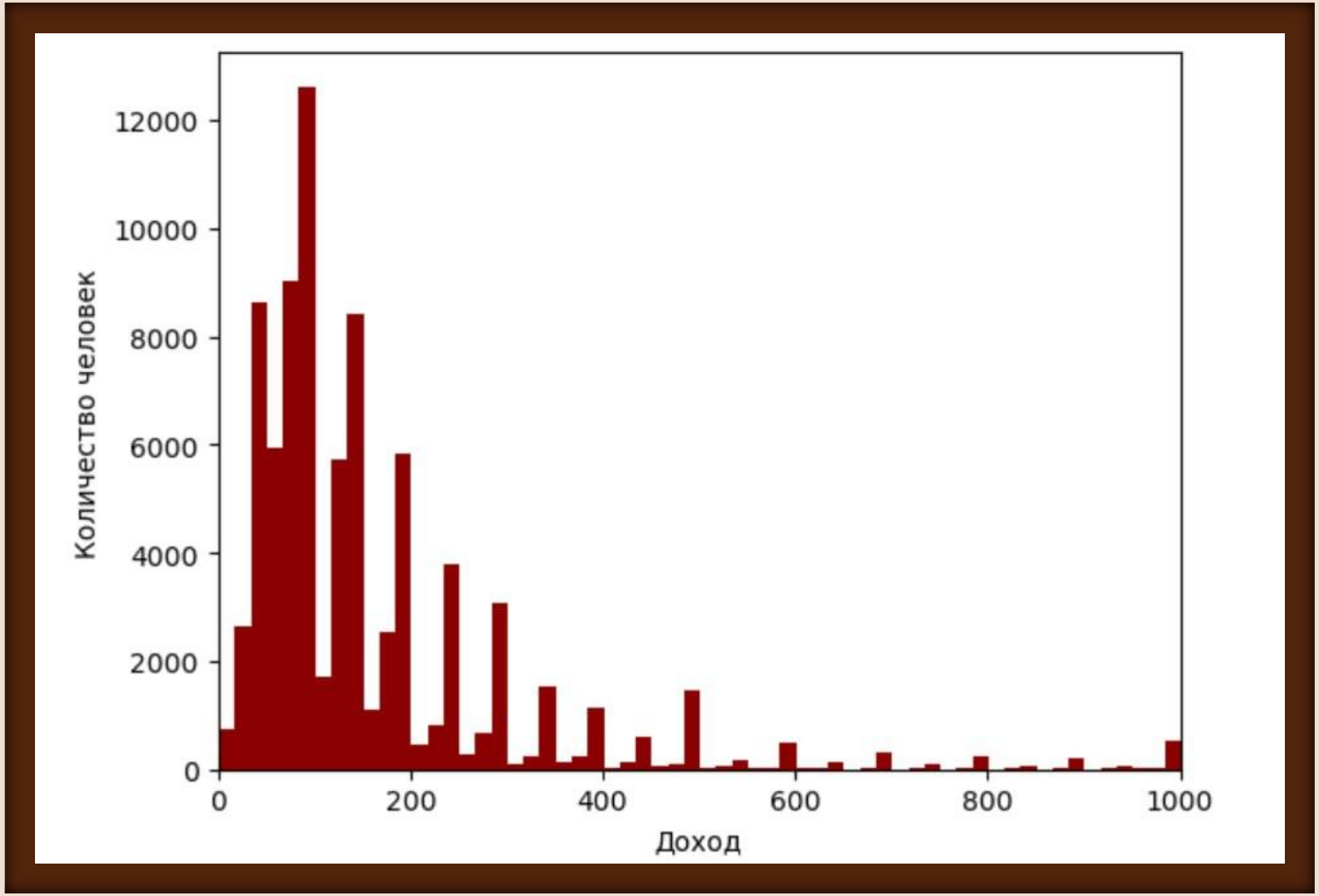
людей, у которых зарплата больше 1000 у.е.

13696

людей, которые были исключены по отсутствию/некорректности данных

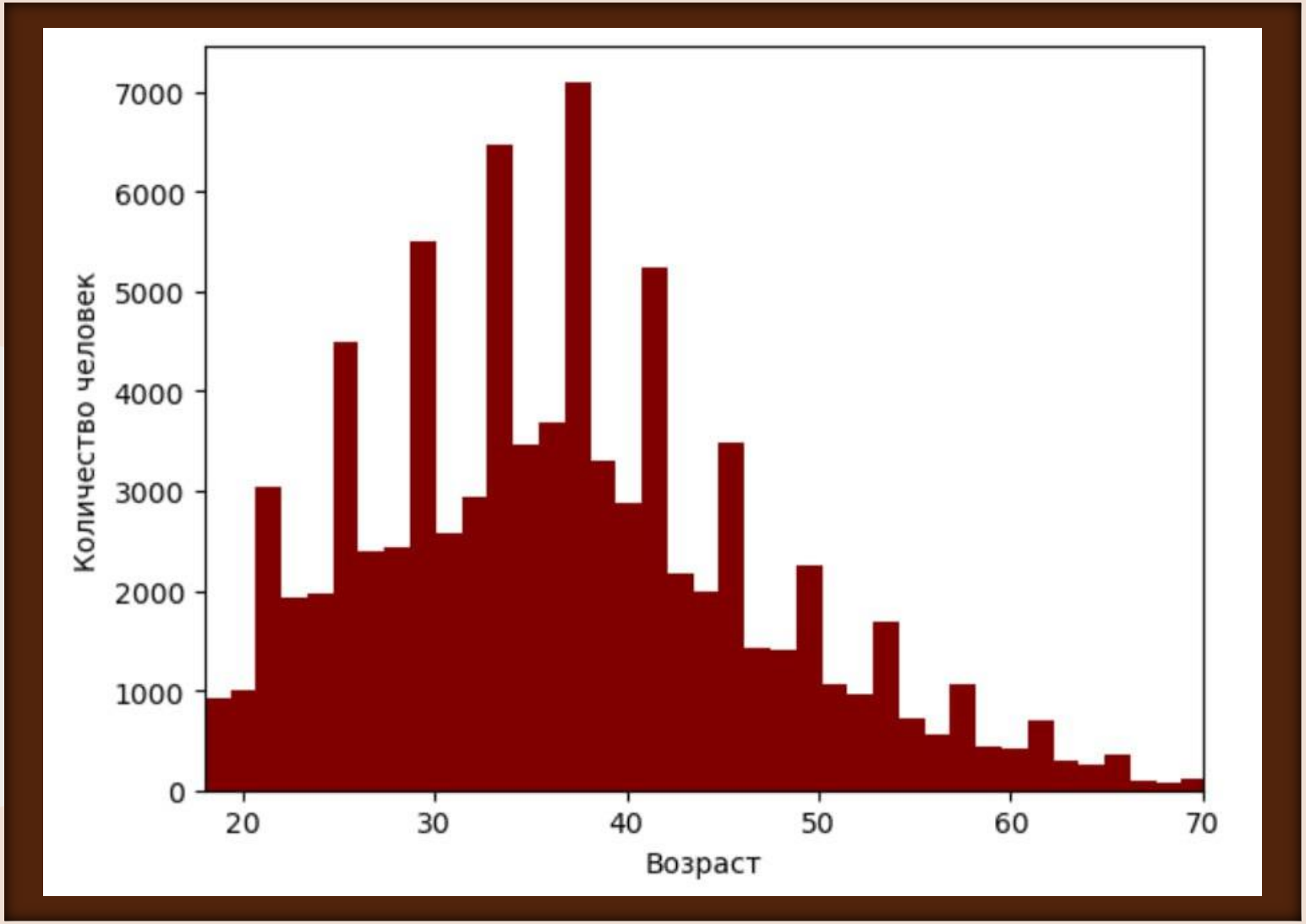
Предварительный анализ

Объем выборки:
82415

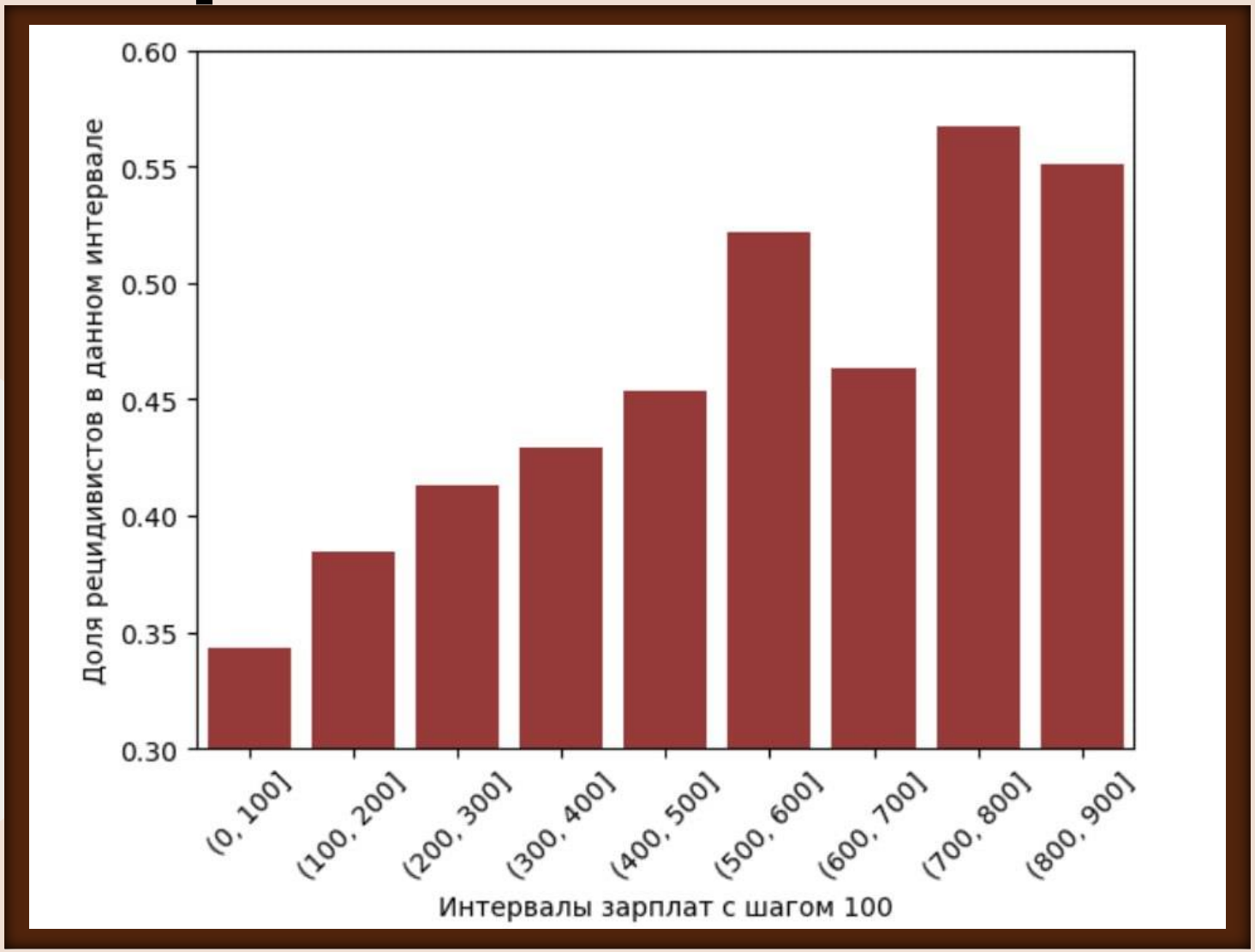


Предварительный анализ

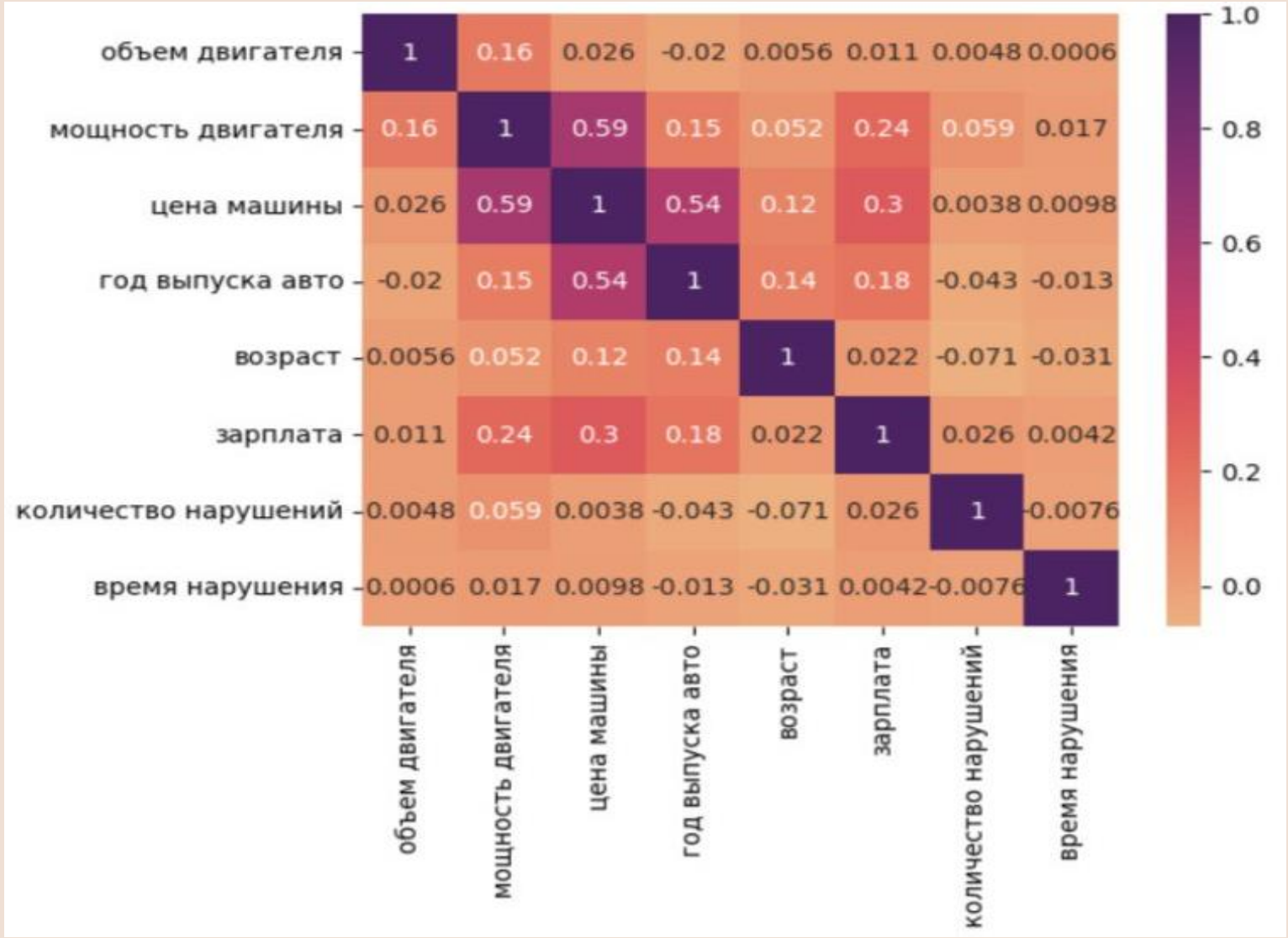
Предварительный анализ



Предварительный анализ



Матрица корреляции



Математическая модель

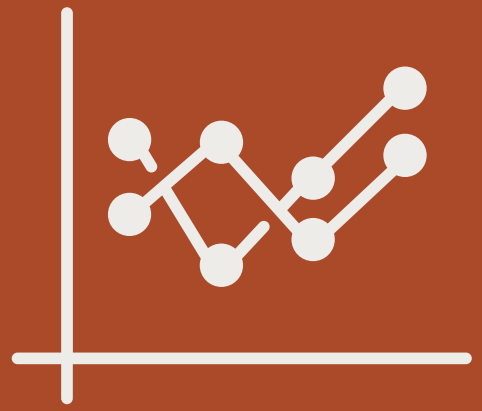
H0 Доли рецидивистов среди богатых и бедных совпадают

H1 Доля рецидивистов среди богатых больше, чем среди бедных

Статистическая значимость 0,01

Метод

Хи-квадрат тест на независимость



Ввод новых переменных

Определим богатых и бедных:

N – граница по зарплате (в условных единицах)

Бедные: зарплата меньше N

Богатые: зарплата больше N

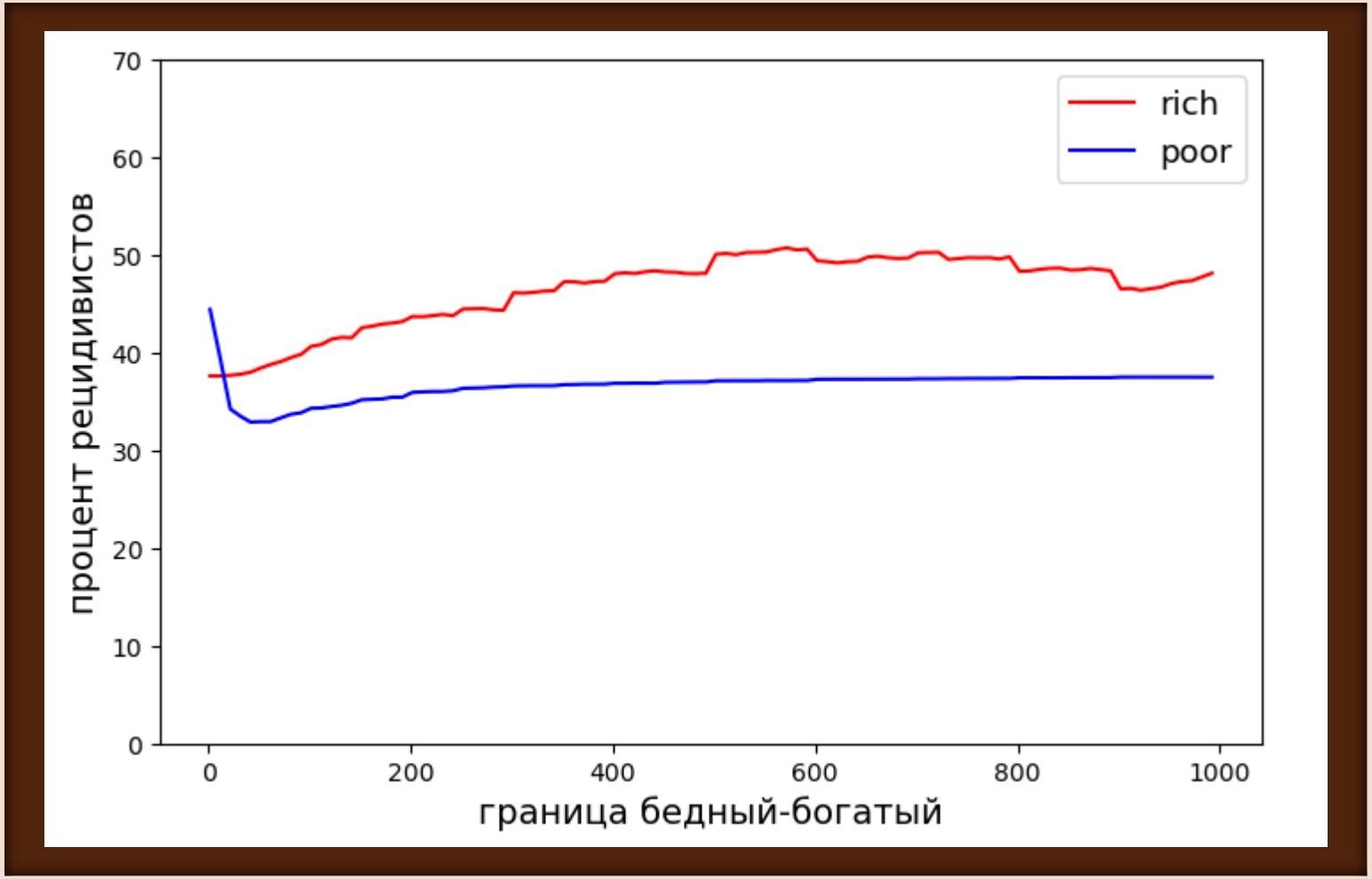
Переменная flag, которая:

= 0, если человек совершил только одно нарушение

= 1, если больше

Введем долю рецидивистов среди богатых и бедных

Общая выборка



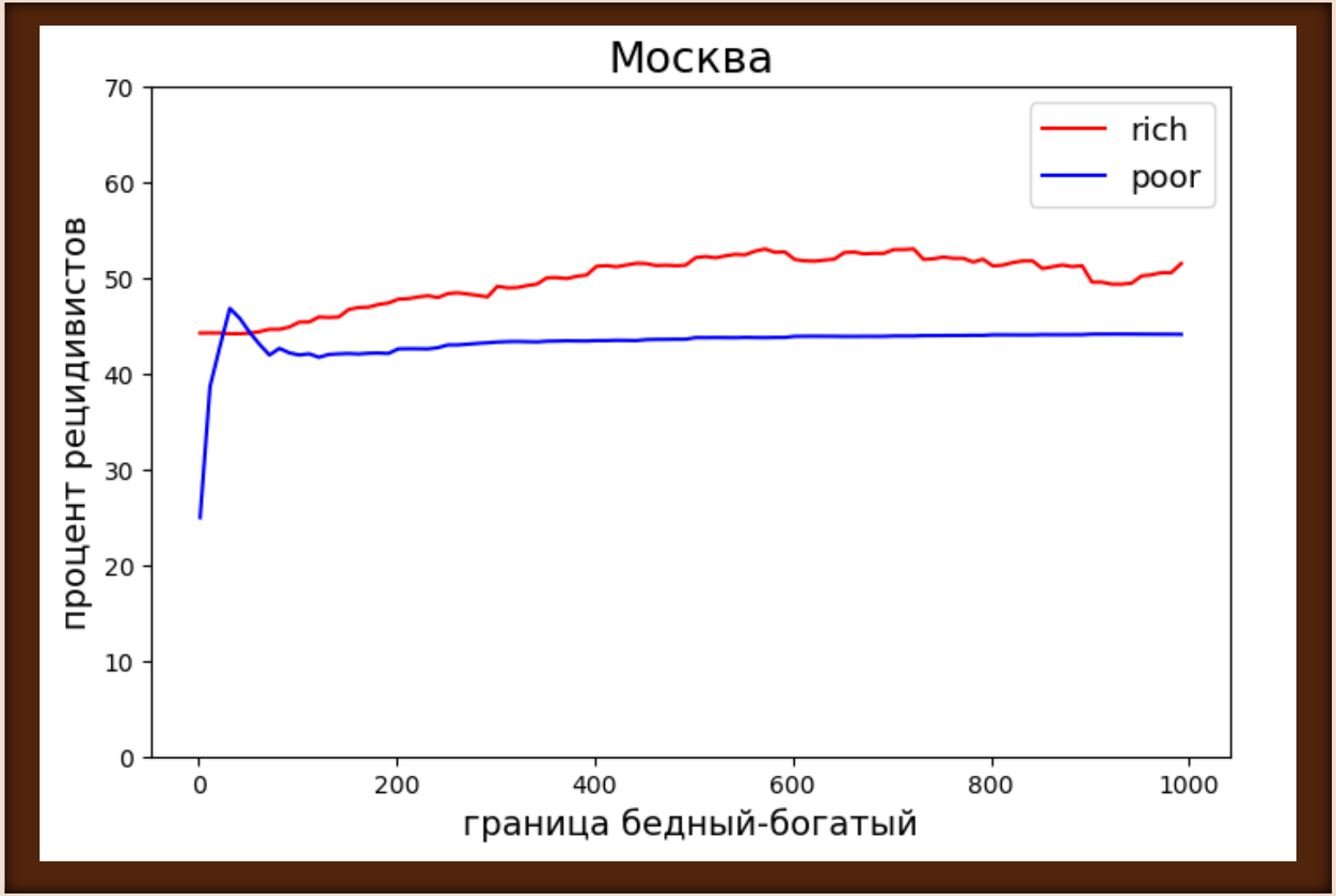
pvalue = 2.7e-57 при N =300

Проверка модели на устойчивость

Алгоритм:

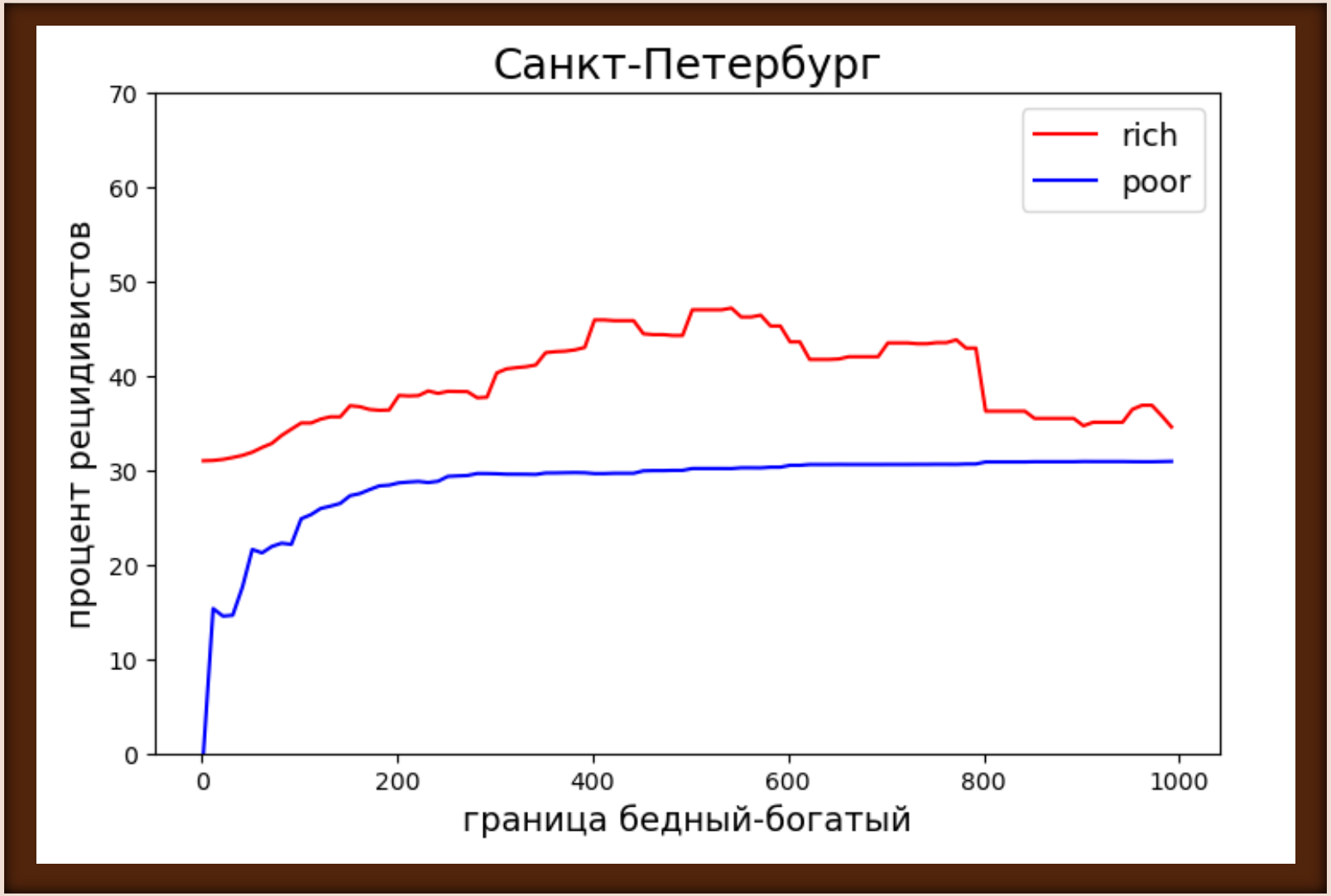
- Разделяем датасет на подвыборки
- Строим графики
- Анализ графиков

Группы по региону



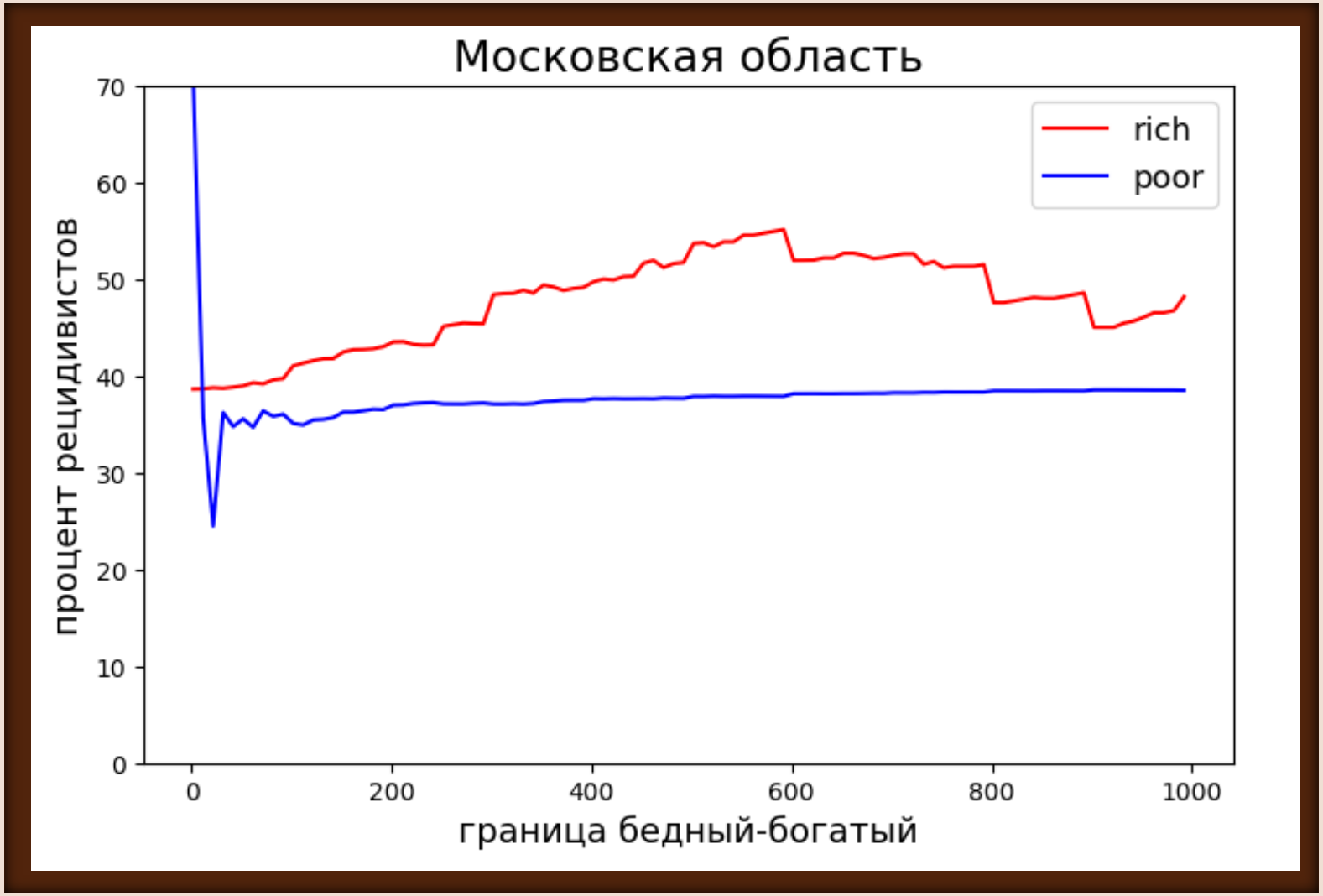
pvalue = 3.2e-08

Группы по региону



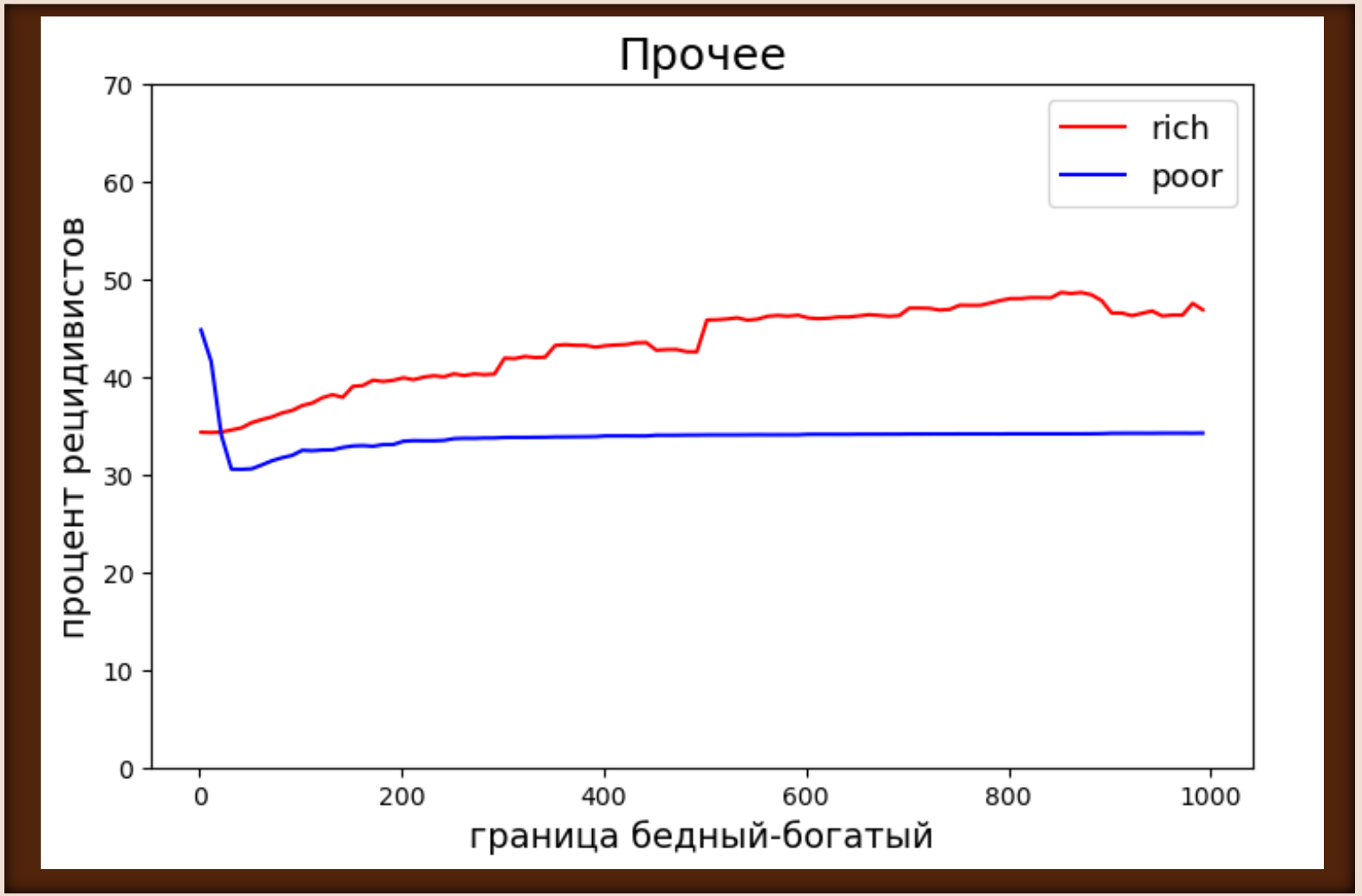
pvalue = 7.2e-05

Группы по региону



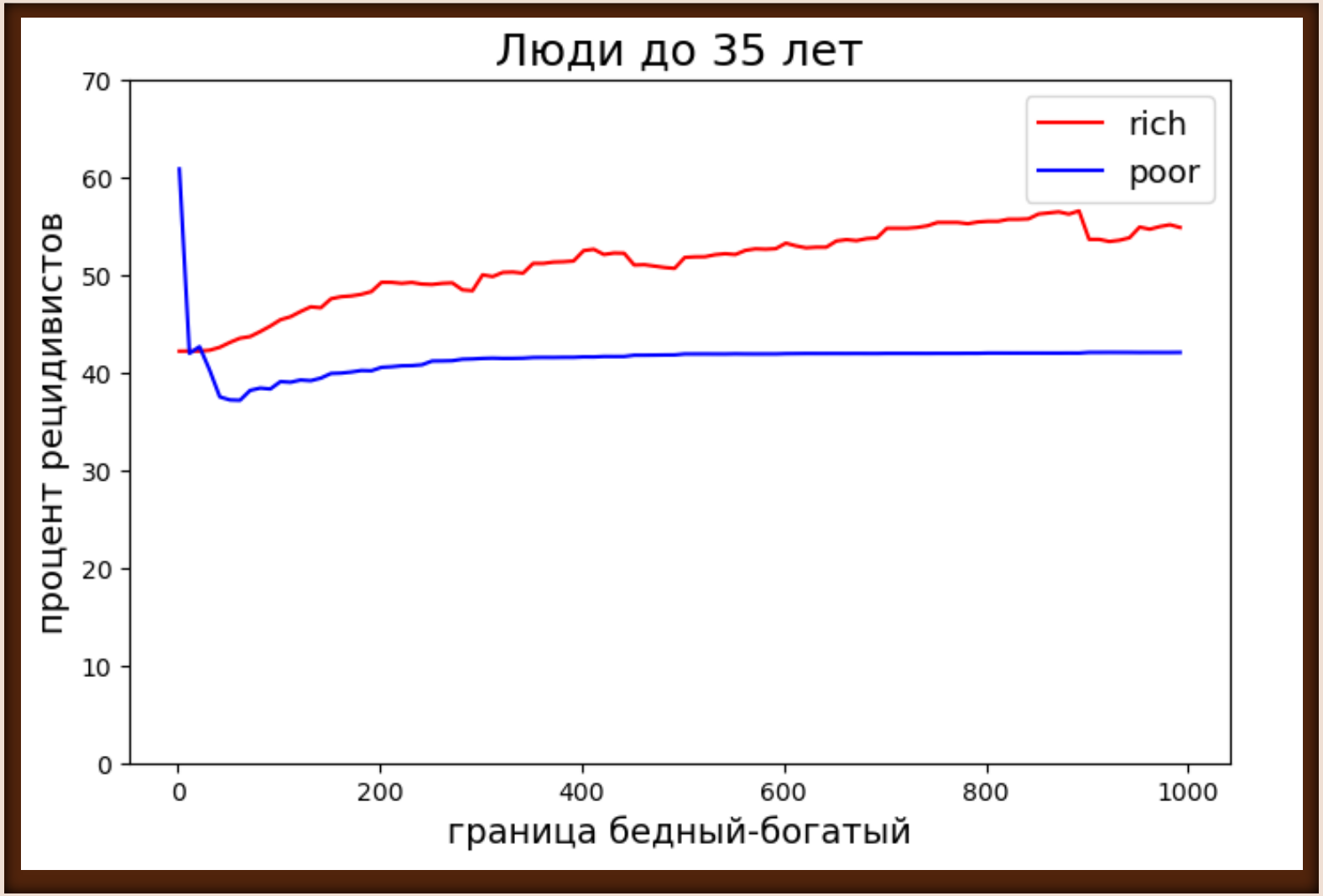
pvalue = 2.0e-11

Группы по региону



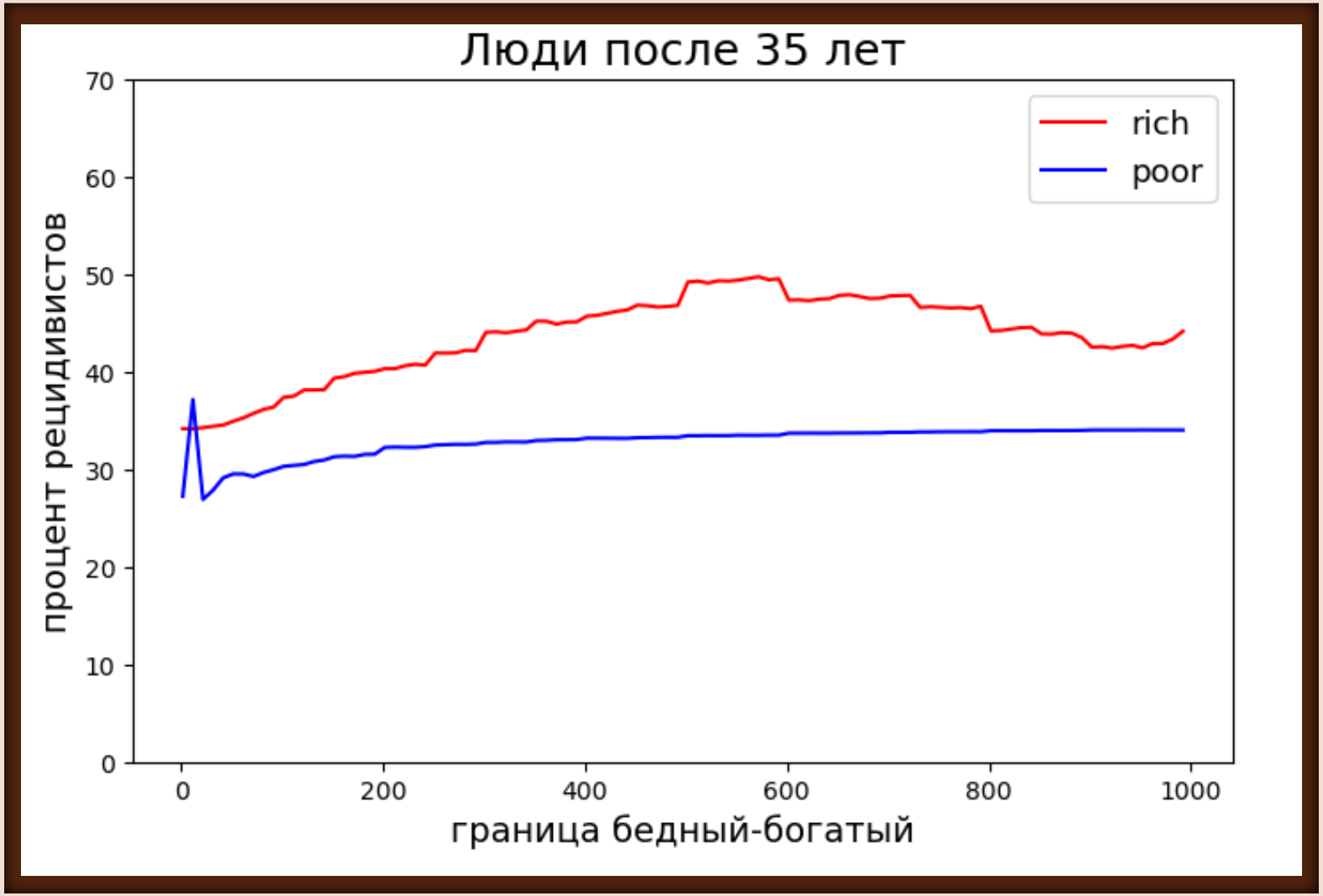
pvalue = 1.5e-13

Группы по возрасту



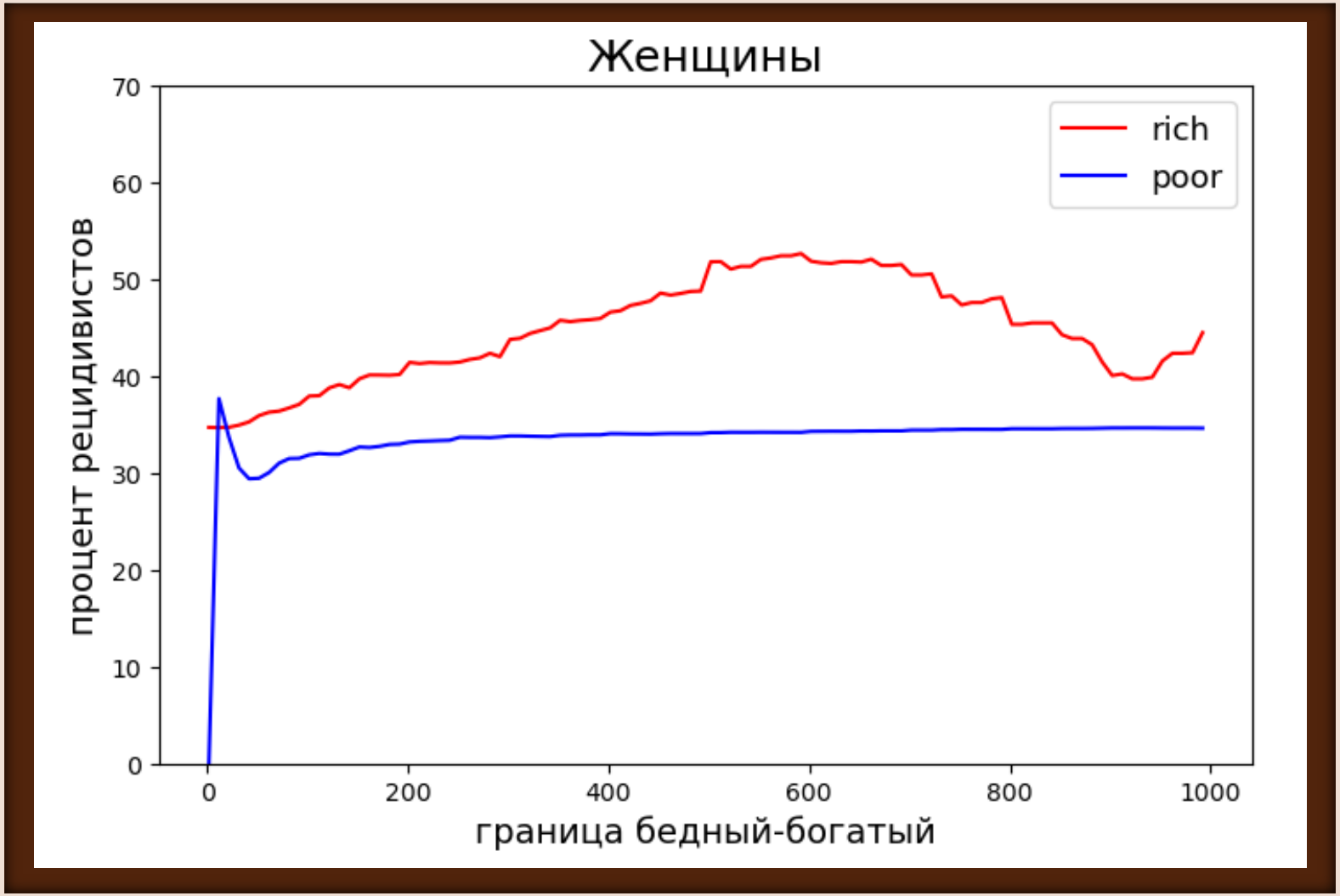
pvalue = 2.8e-16

Группы по возрасту



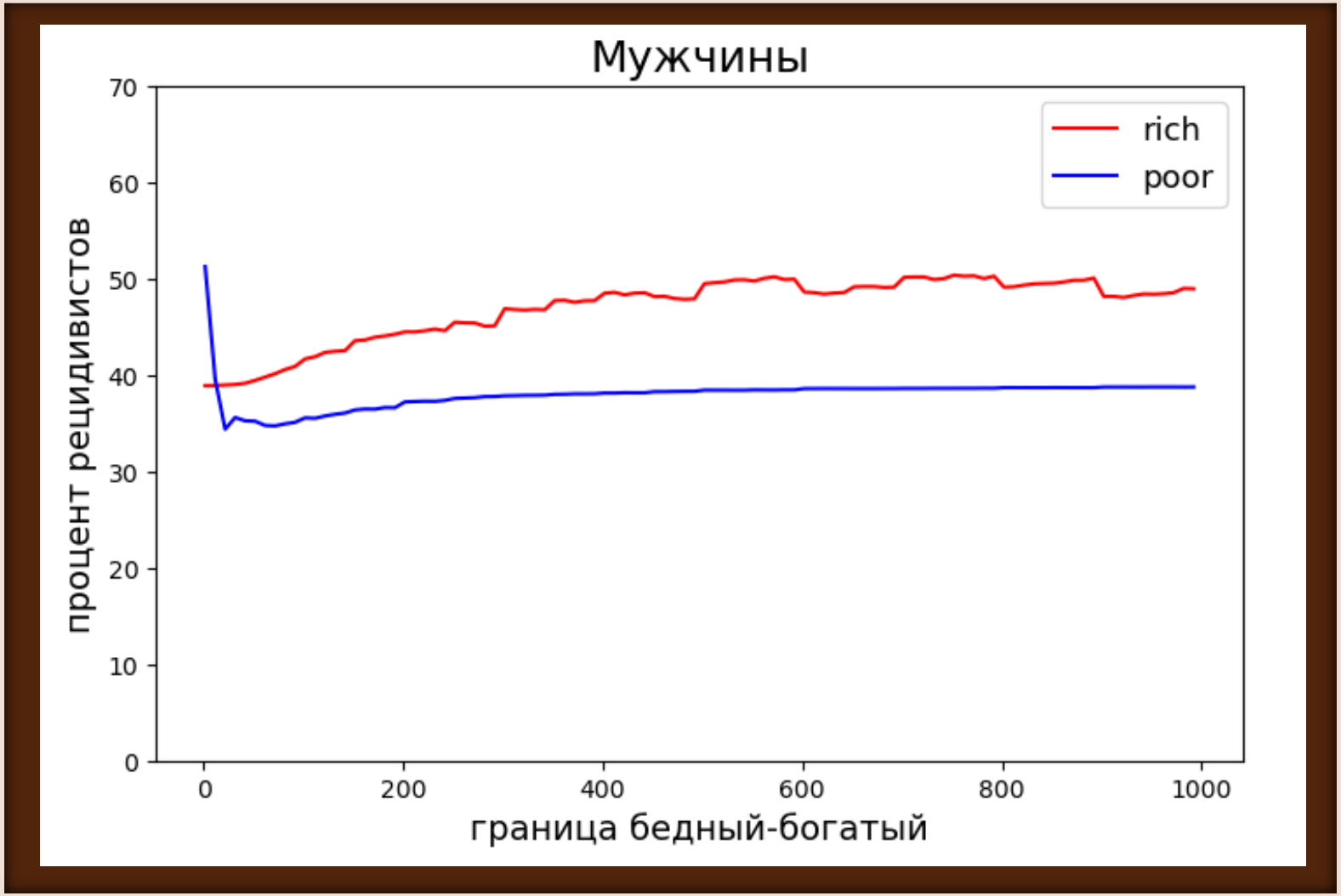
pvalue = 4.1e-56

Группы по полу



pvalue = 9.6e-18

Группы по полу



pvalue = 4.9e-36

Гипотеза подтвердилась

и оказалась устойчивой на всех подвыборках

Доля рецидивистов среди богатых больше, чем среди бедных

Ограничения и перспективы

- Только пользователи Т-Банка
- Ограниченный период времени
- Неполная выборка
- Не учитывается сезонность

- Использование данных из полного реестра штрафов по РФ
- Использовать целиком данные нескольких лет
- Учесть внешние факторы
- Оптимизировать систему штрафов, чтобы убрать различие между отношениями "богатых" и "бедных" к штрафам

Альтернативные механизмы

богатые имеют более мощные машины,
поэтому они чаще нарушают правила,
которые ограничивают их вождение

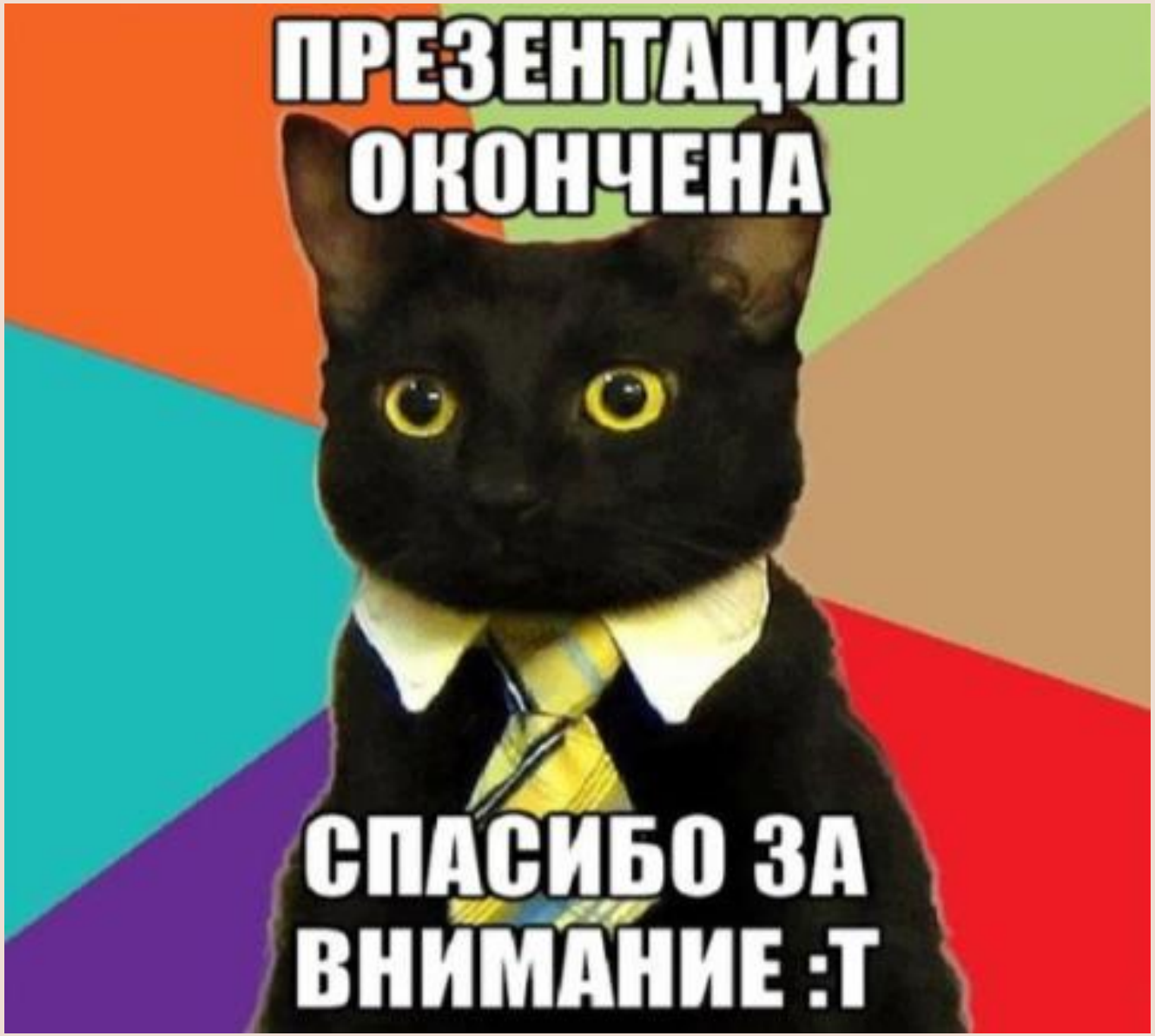
Крутая бизнес идея

Что хотим сделать?

оптимизировать систему штрафов, чтобы убрать различие между отношениями "богатых" и "бедных" к штрафам

Как можно реализовать?

обновление сбора штрафов (система поощрения для получающих меньшее количество штрафов и санкции для часто получающих)



Спасибо

Хи-квадрат тест на независимость

$$x^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i}$$

Mi – ожидаемые величины

	богатые	бедные	общее число
рецидивисты	a	b	A+b
не рецидивисты	C	d	C + d
общее число	A+c	B+d	A+B+c+d

Например, m1 = (a+c)(a+b)/(a+b+c+d)
Xi- наблюдаемые величины
DF = (r-1)(c-1)
R – количество строк
C – количество столбцов
Df – количество степеней свободы

Первая гипотеза

Люди с маленькой зарплатой чаще нарушают утром, чем вечером

Анализ выборки