



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

ФЛЕШБЕКИ И ПРОБЛЕМЫ НА ЭТАПЕ ПОДГОТОВКИ ДАННЫХ К ОБУЧЕНИЮ

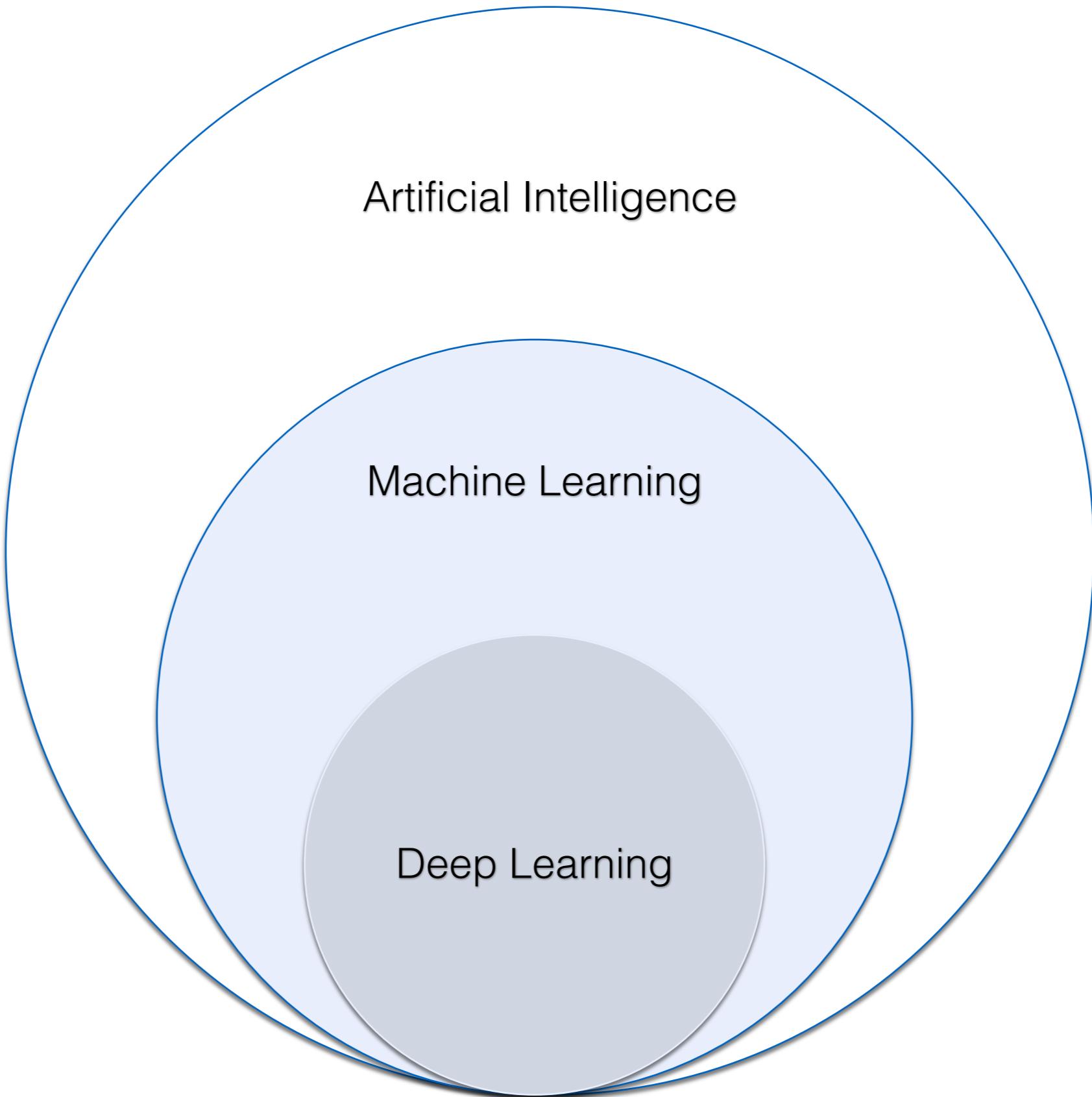
Теванян Элен

12.11.2019

Москва 2019



FLASHBACK



ТЕРМИНОЛОГИЯ

- x (**sample**) – объект, для которой хотим делать предсказания
 - Поездки
 - y (**target**) – ответ, целевая переменная, т.е. То, что хотим предсказать
 - Длительность поездки
-
- $(x_i, y_i)_{i=1}^{\ell}$ – обучающая выборка, прецеденты, т.е. все объекты, для которых известны значения целевого признака
 - ℓ – размер выборки.

ПРИЗНАКИ

- Компьютер умеет работать с числовой информацией
- Объекты характеризуются числовой информацией – признаками, факторами, «фичами» (от англ. features)
- m – число признаков
- $x = (x^1, \dots, x^m)$

AI, ML, DL

- Искусственный интеллект – широкая область, в которой изучают процесс принятия решений.
- Машинное обучение – подобласть искусственного интеллекта, в которой на основании данных машины учатся принимать решения без прямого, явного программирования по сценариям.
- Глубокое обучение – подобласть машинного обучения, сфокусированная на нейронных сетях.

ОБУЧЕНИЕ

- $a(x)$ – алгоритм/модель
- Это функция, предсказывающая ответ для любого объекта
- Функция потерь (ошибок) – мера корректности алгоритма
- Для задачи регрессии можно использовать среднеквадратическую ошибку Mean Square Error:

$$MSE = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

ОБУЧЕНИЕ

- Функция потерь – один из важнейших компонентов при анализе данных и должна соответствовать бизнес требованиям.

ОБУЧЕНИЕ

- Функция потерь – один из важнейших компонентов при анализе данных и должна соответствовать бизнес требованиям.
- Есть прецеденты
- Определен функционал качества
- Есть параметризованное семейство алгоритмов:
«Если время после α часов, то длительность заказа сокращается на 10%»

ОБУЧЕНИЕ

- Функция потерь – один из важнейших компонентов при анализе данных и должна соответствовать бизнес требованиям.
- Есть прецеденты
- Определен функционал качества
- Есть параметризованное семейство алгоритмов:
«Если время после α часов, то длительность заказа сокращается на 10%»

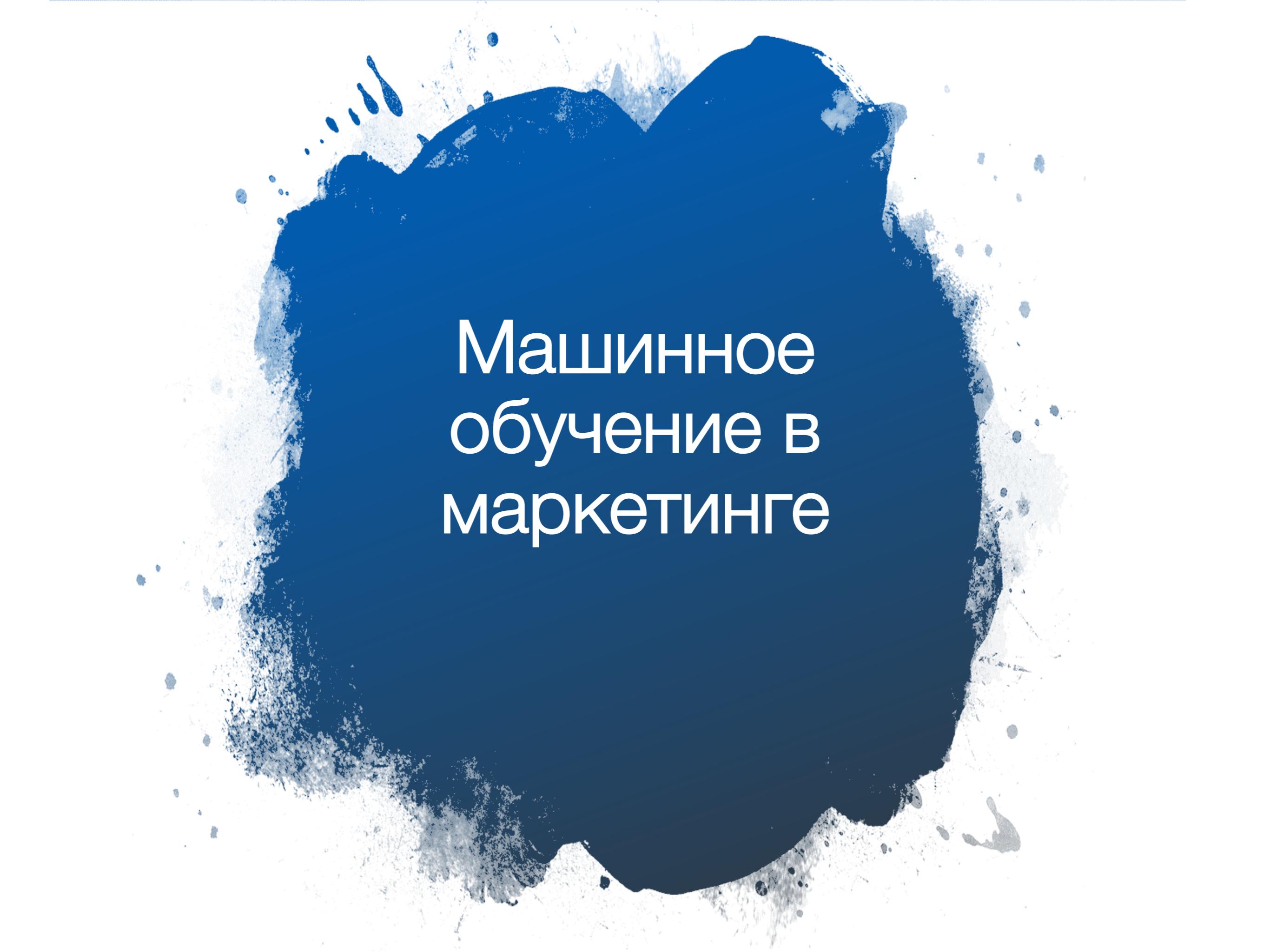
Обучение – поиск оптимальных алгоритмов с точки зрения функционала качества.

ПРЕДСКАЗАНИЕ ЦЕНЫ НА ТОВАР

- Задача: товар → цена
- x_i – объект, для которого строим предсказания (i -ый товар)
- y_i – целевая переменная (цена на i -ый товар)
- (x_i, y_i) – прецедент
- Обучающая выборка – набор всех прецедентов

Как решить эту задачу?

Найти алгоритм $a(x)$: $a(x_i) \approx y_i$



Машинное обучение в маркетинге

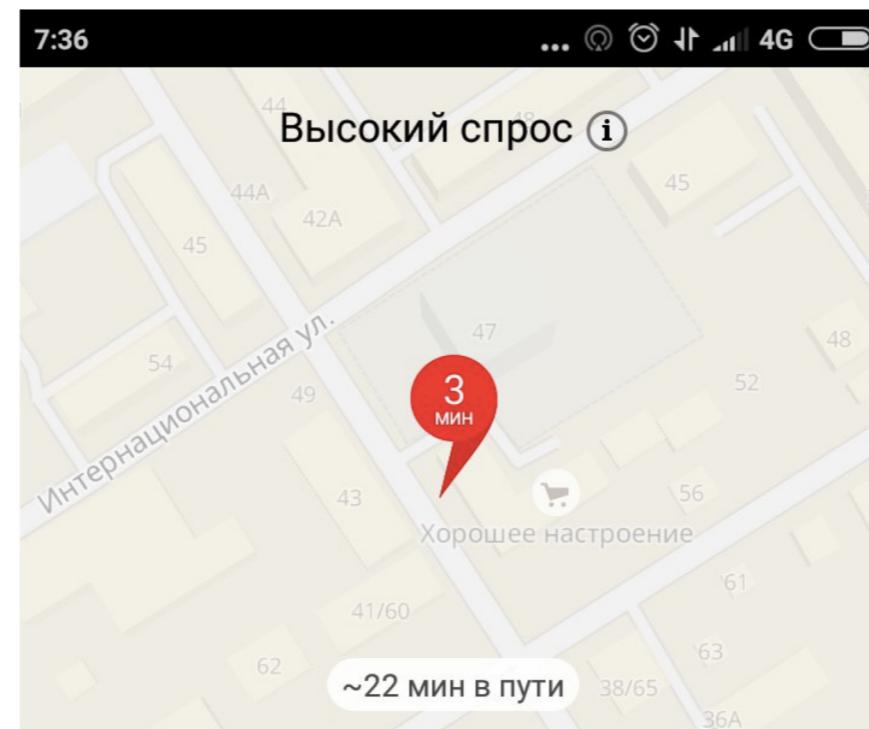
МАРКЕТИНГОВАЯ АНАЛИТИКА

- Descriptive
- Predictive
- Prescriptive

СПРОС



ЦЕНООБРАЗОВАНИЕ



📍 Максима Горького, 38А/58 Подъезд

📍 Шумакова, 74Б +



ЭКОНОМ

425₽



КОМФОРТ

286₽

Комментарий,
пожелания

Способ оплаты
VISA 9296

Вызвать такси

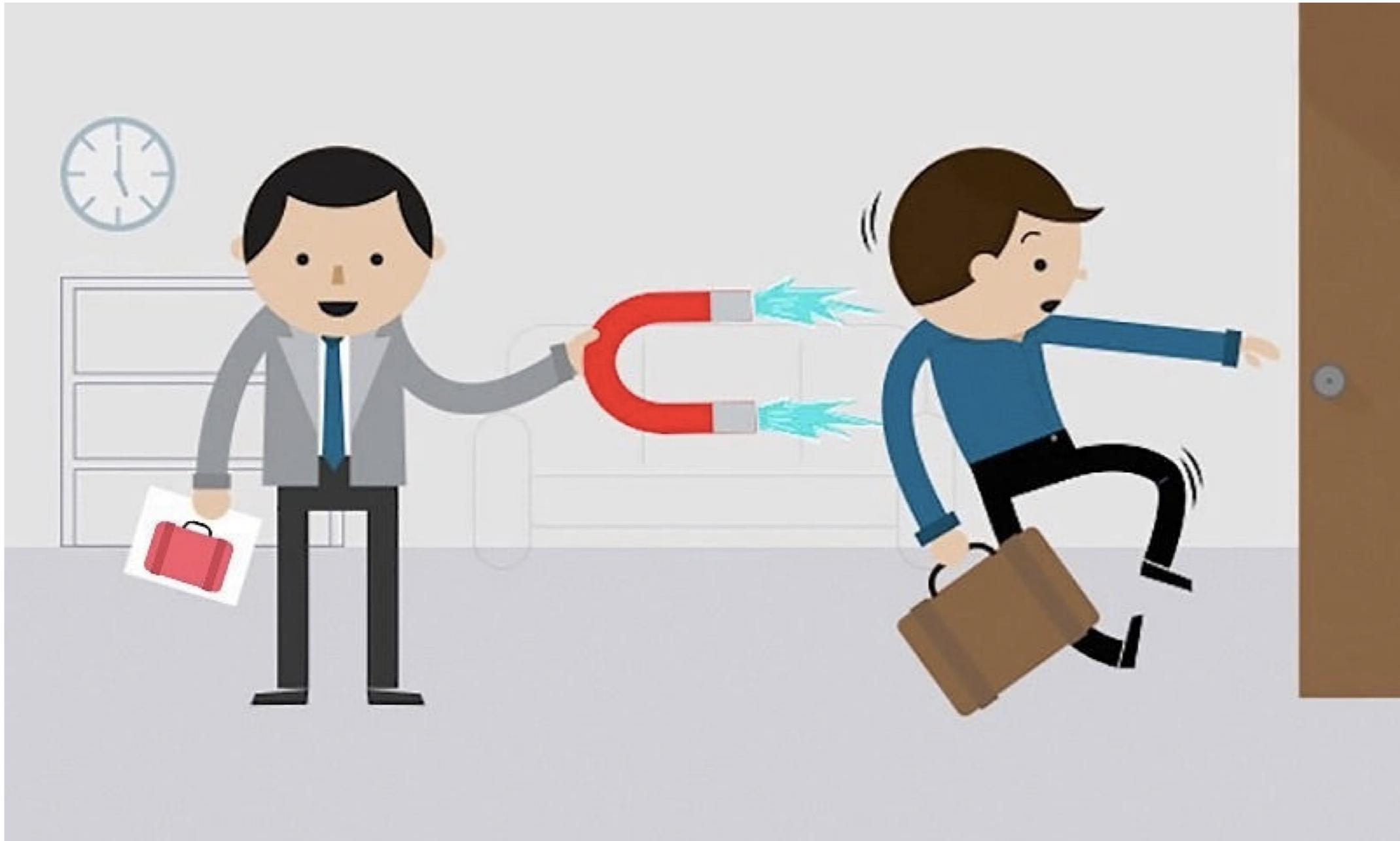
ПРОМО-КАМПАНИИ



будь
крутым кабаном —
найди трюфель



ОТТОКИ КЛИЕНТОВ



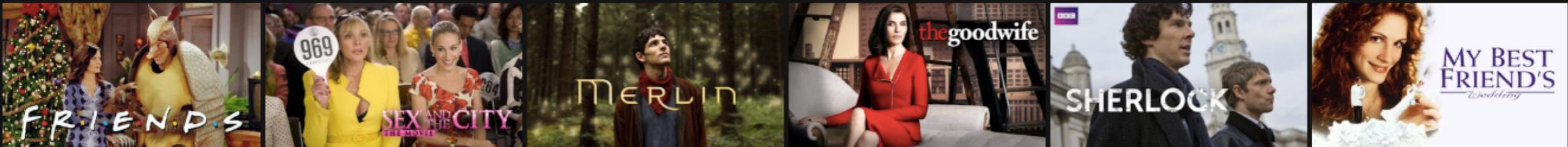
РЕКОМЕНДАЦИИ КЛИЕНТАМ



Home TV Shows Movies Recently Added My List

SEARCH KIDS BELL SMILEY

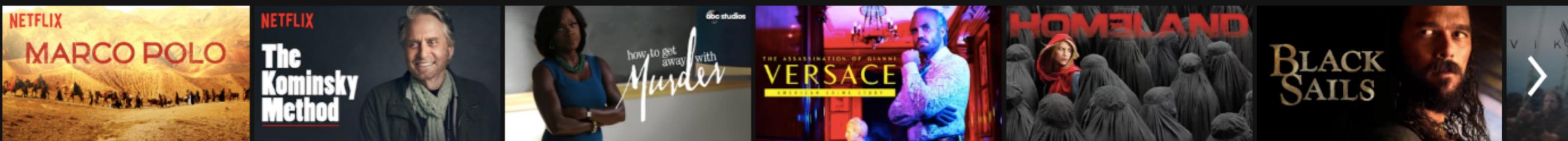
Watch It Again



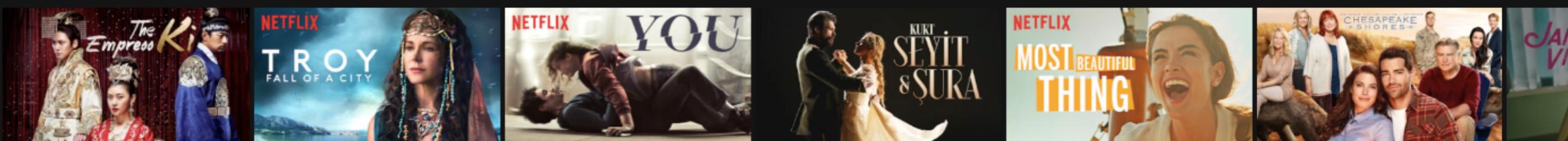
Because you watched Outlander



Critically-acclaimed TV Dramas >



Romantic TV Dramas



СЕГМЕНТАЦИЯ КЛИЕНТОВ





Как сделать ML в
маркетинге и не
умереть

ЭТАПЫ МАШИННОГО ОБУЧЕНИЯ

- Сбор данных
- Подготовка данных
- Обучение модели и валидация
- Анализ эффективности модели

СБОР ДАННЫХ

- Самый тяжелый этап
 - Внутренние данные

Все, что находится внутри компании
 - Внешние данные

Соцсети
Коллaborации с другими компаниями
Биржи данных

СБОР ДАННЫХ

- Самый тяжелый этап
 - Внутренние данные

Все, что находится внутри компании
 - Внешние данные

Соцсети
Коллaborации с другими компаниями
Биржи данных
- Мы на курсе не занимаемся сбором данных, но вы теперь питонисты ☺

ПОДГОТОВКА ДАННЫХ

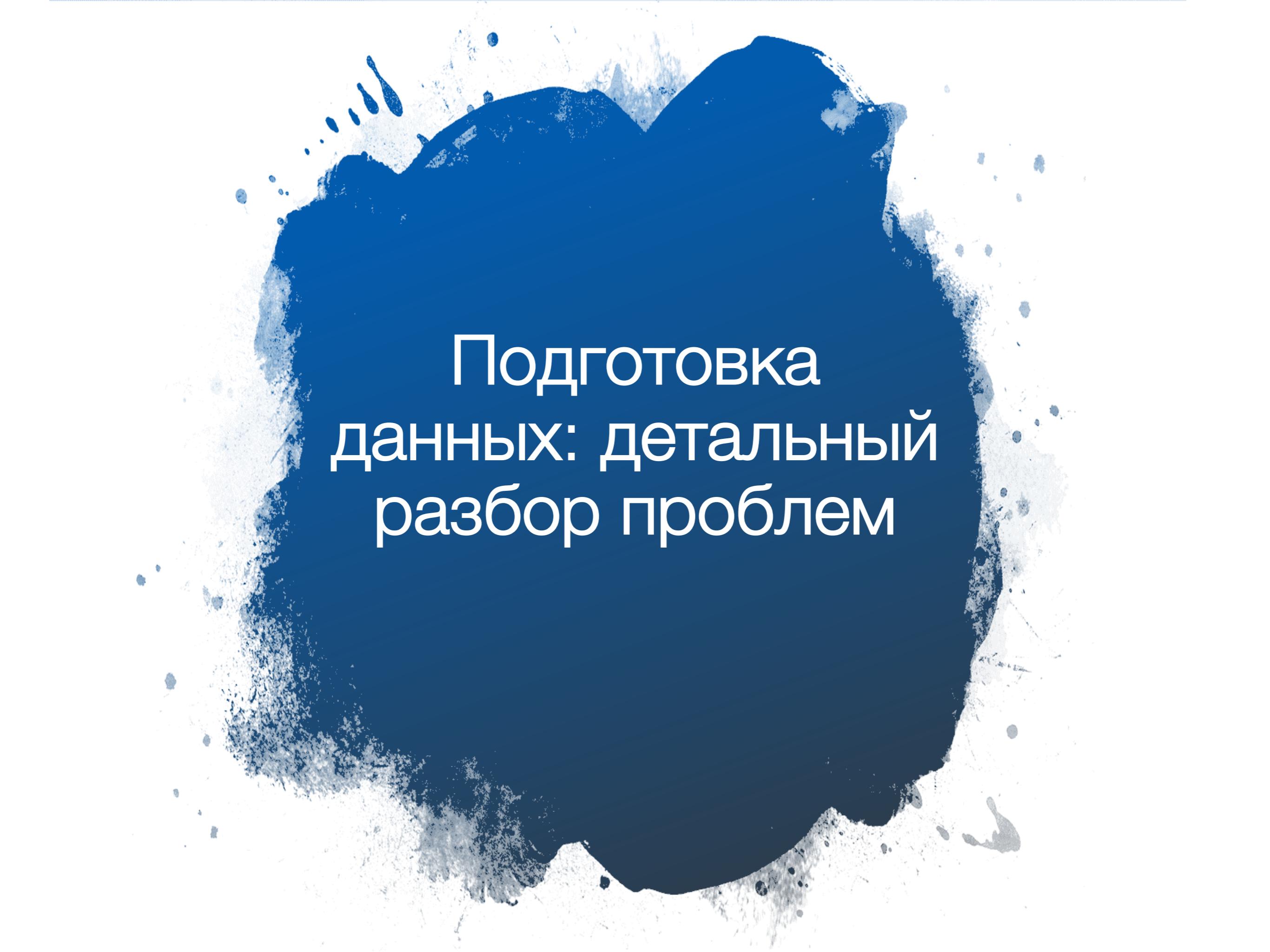
- Заполучить сырье данные – это не решение всех проблем
 - Проблемы с целевой переменной
 - Проблемы с числовыми переменными
 - Проблемы с категориальными переменными
 - Проблемы с количеством переменных
- В общем, все проблема

ОБУЧЕНИЕ МОДЕЛИ И ВАЛИДАЦИЯ

- Применить подходящие модели
- Подобрать параметры (тюнинг модели)
- Кросс-валидацией убедиться в качестве модели

АНАЛИЗ ЭФФЕКТИВНОСТИ МОДЕЛИ

- Живой эксперимент
 - АБ-тестирования
 - Байесовские многорукие бандиты



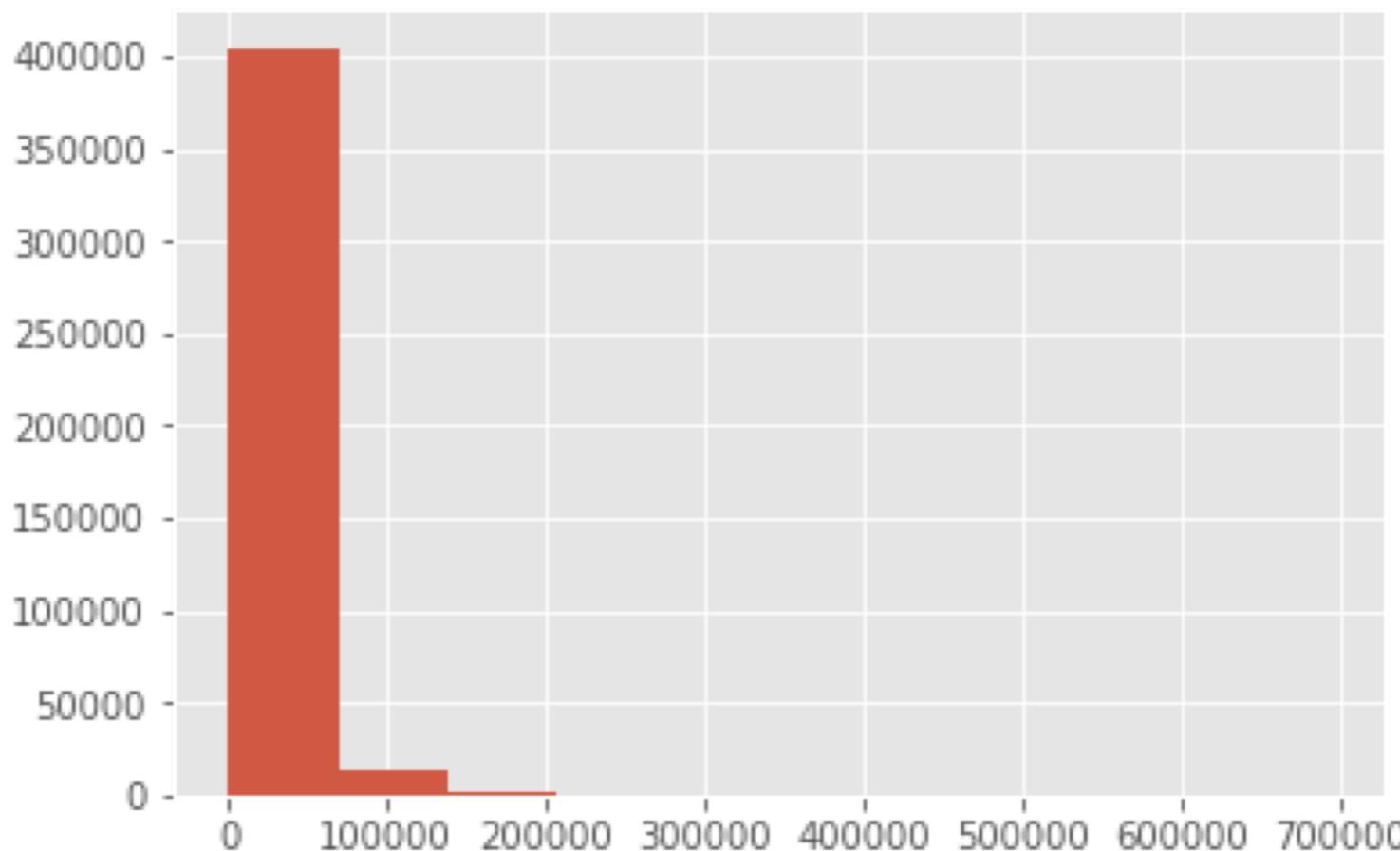
Подготовка данных: детальный разбор проблем

Целевая переменная

Проблемы, которых никто не ждал

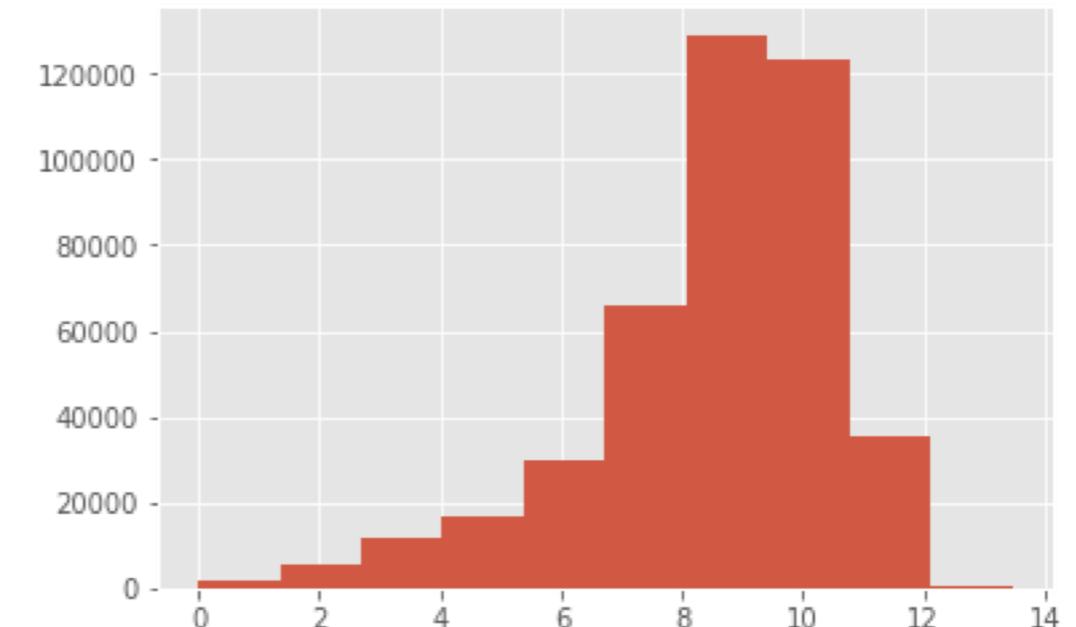
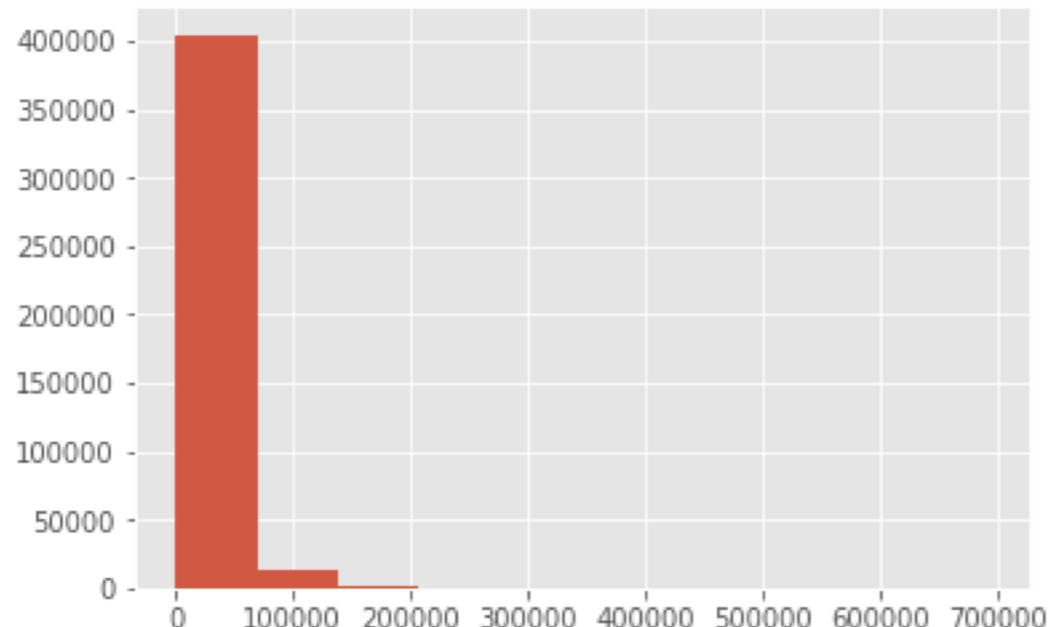
РЕГРЕССИЯ

- ДЛИННЫЙ ХВОСТ



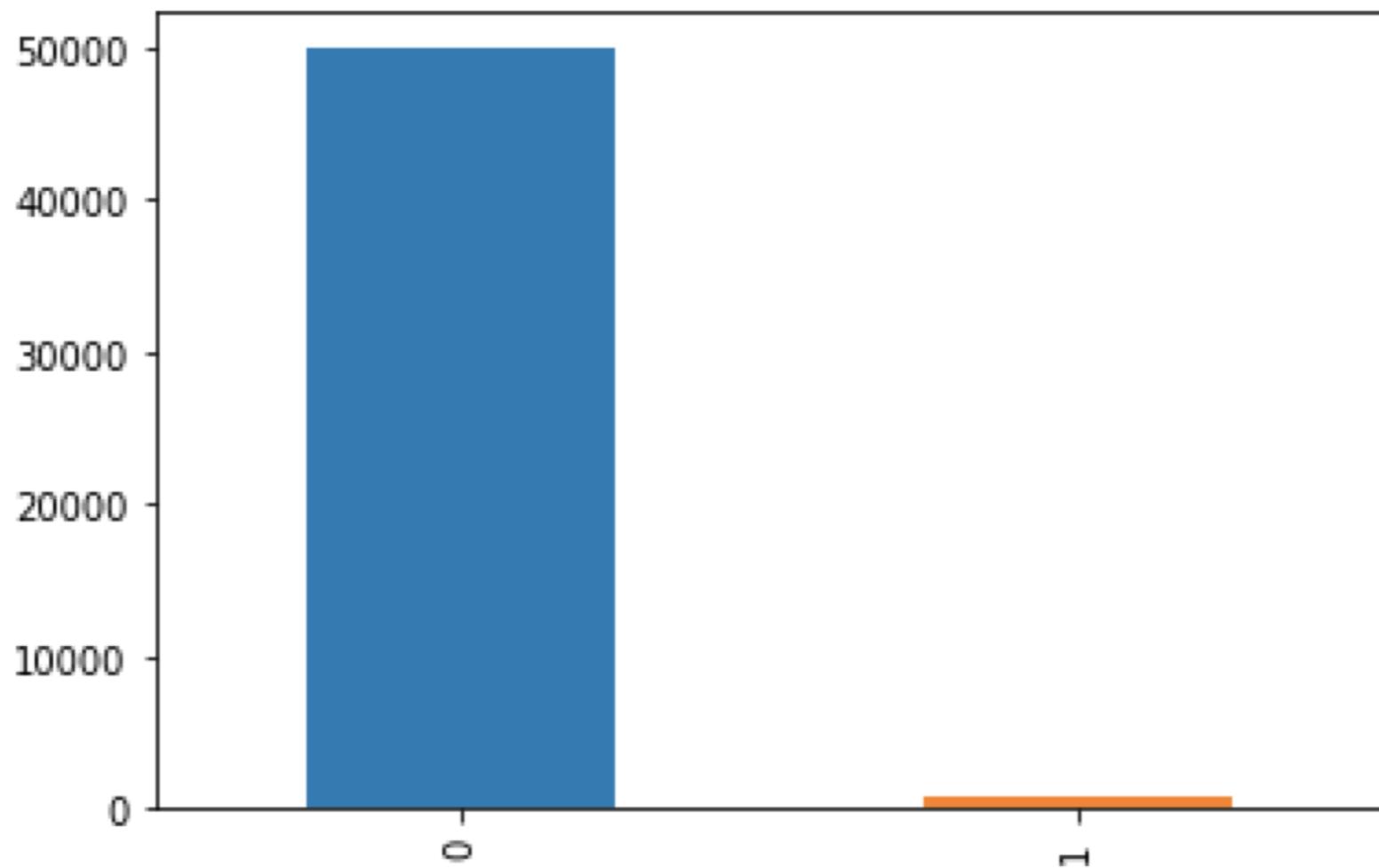
РЕГРЕССИЯ

- Решение: переход в логарифмическую шкалу



КЛАССИФИКАЦИЯ

- Дисбаланс классов



КЛАССИФИКАЦИЯ

- Решение: балансировка классов



UPAMPLING

- Добавление в данных дубликатов редкого класса

SUBSAMPLING

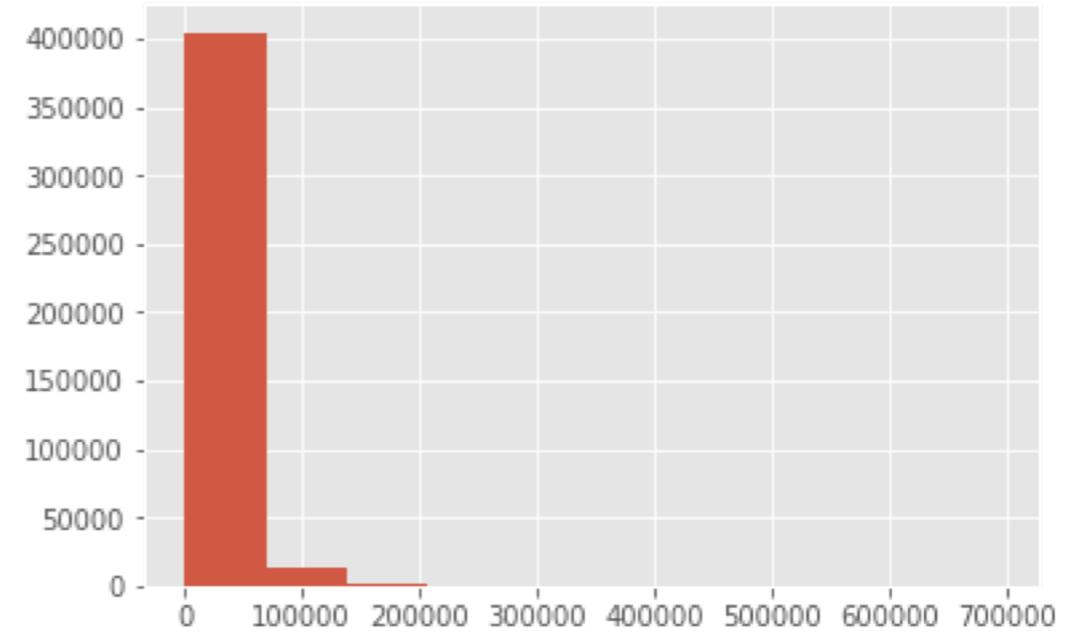
- Генерация подвыборки данных из преобладающего класса

Числовые переменные

Проблемы, которых никто не ждал [2]

ЧТО НЕ ТАК С ЧИСЛОВЫМИ ПЕРЕМЕННЫМИ?

- Длинный хвост



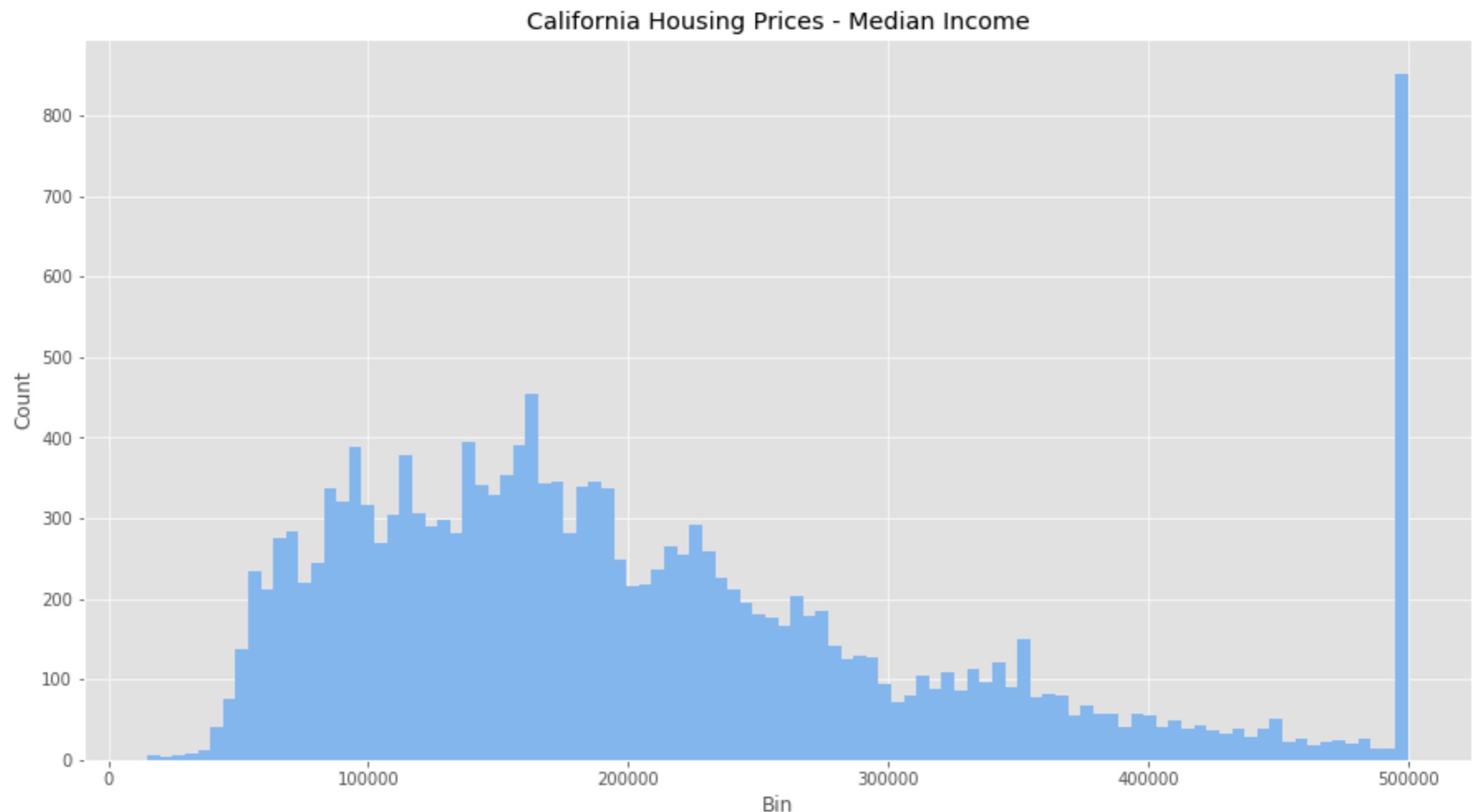
- Разные шкалы измерений

НОРМАЛИЗАЦИЯ

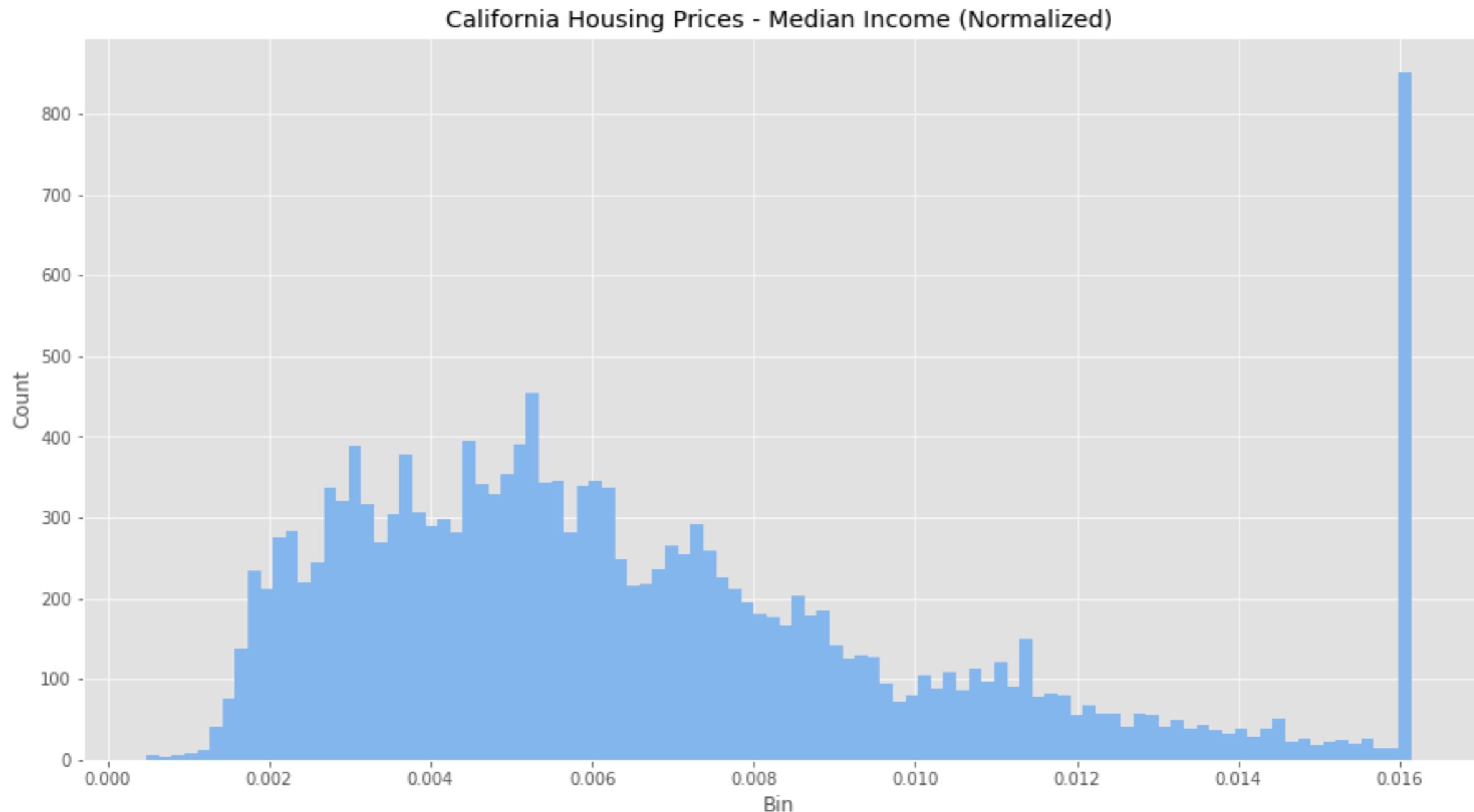
- Замена признаков так, что они лежат в интервале от 0 до 1

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

НОРМАЛИЗАЦИЯ

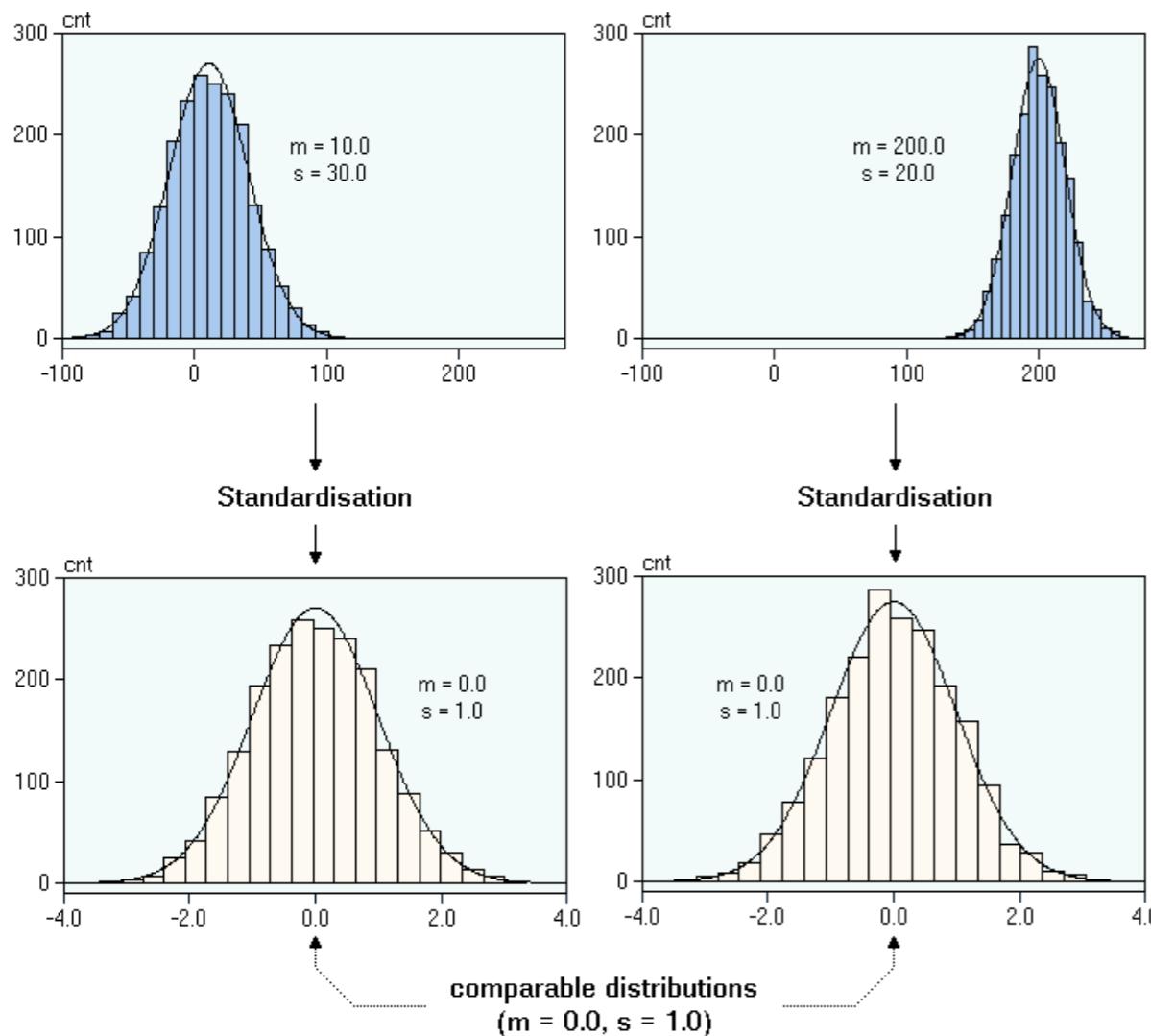


НОРМАЛИЗАЦИЯ



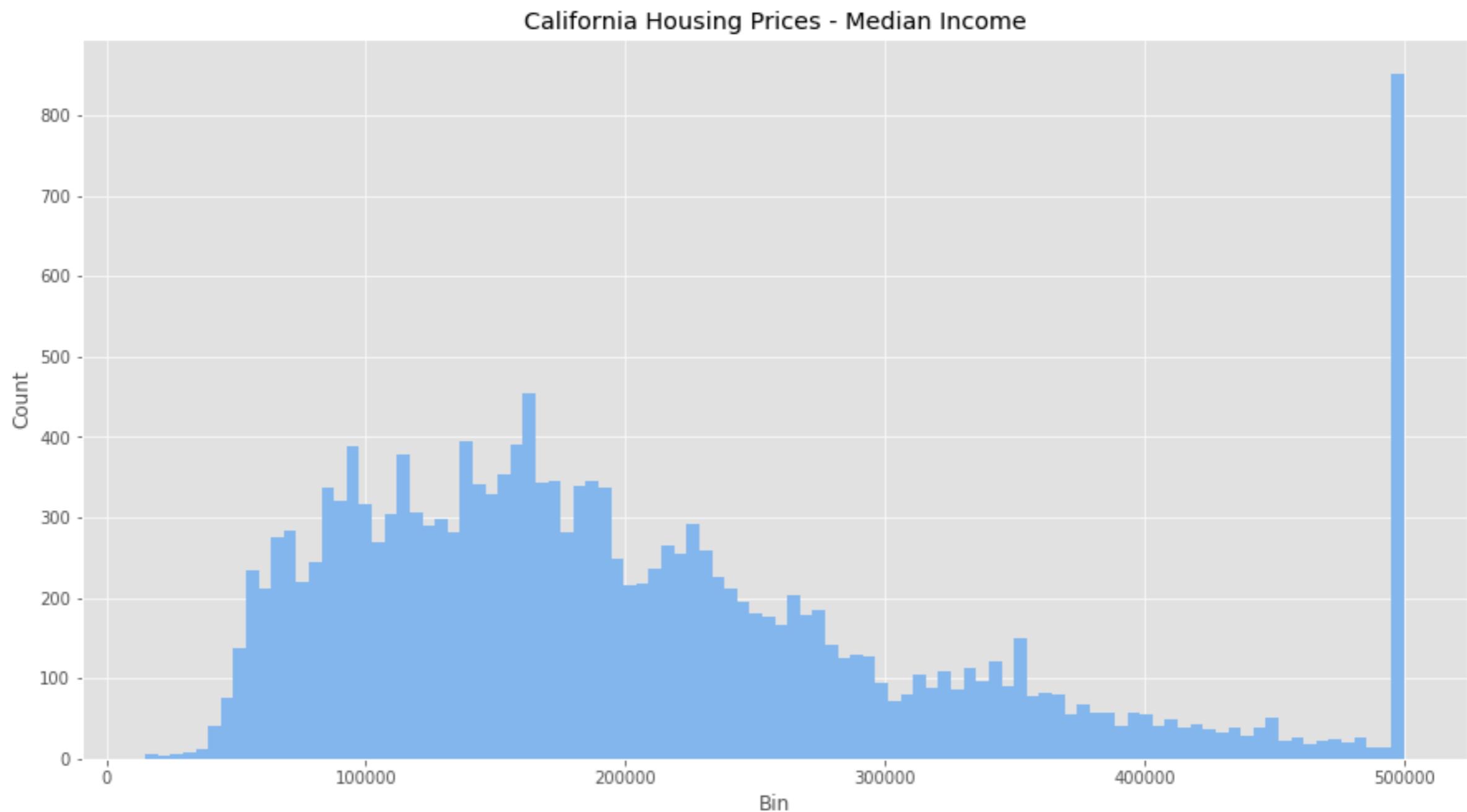
СТАНДАРТИЗАЦИЯ

- Замена признаков так, что:
 - среднее 0
 - дисперсия 1

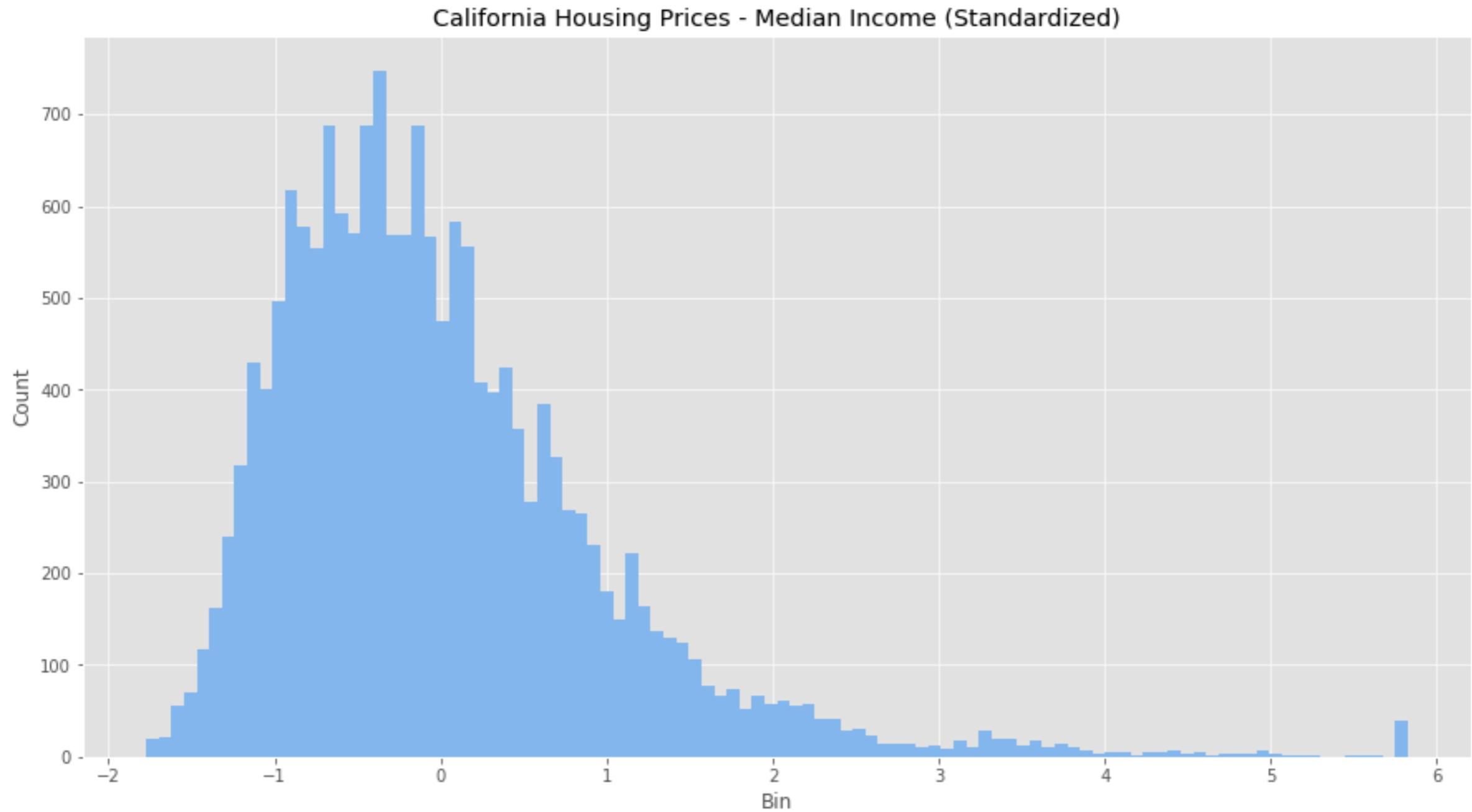


$$z = \frac{x_i - \mu}{\sigma}$$

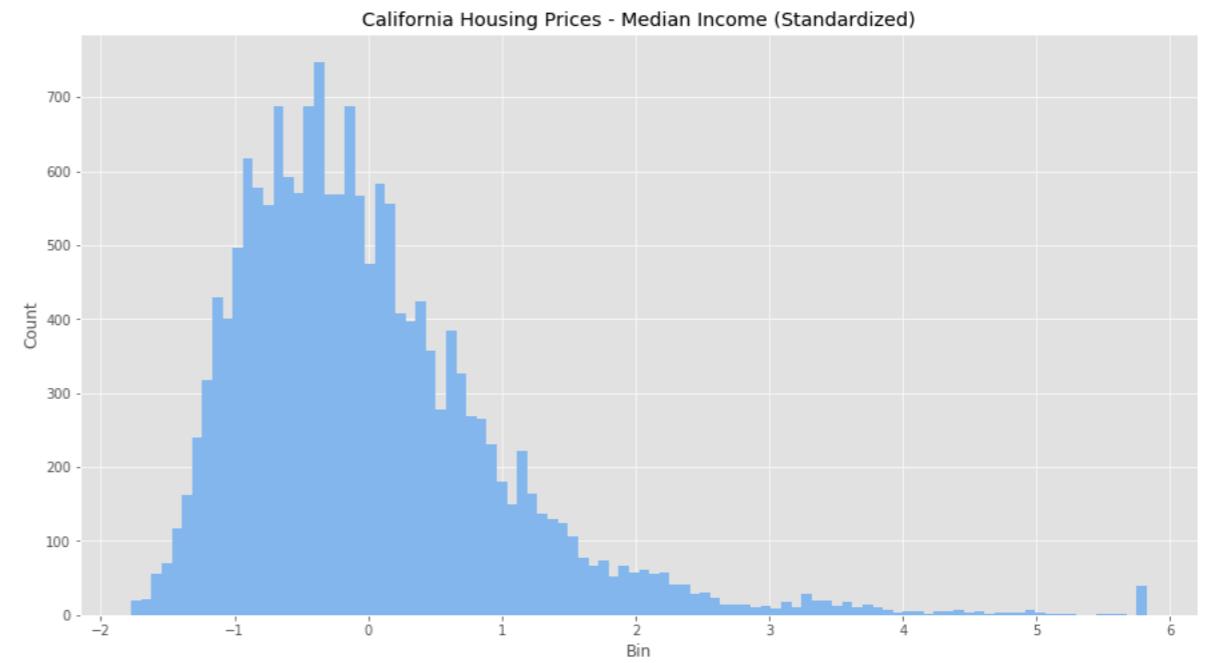
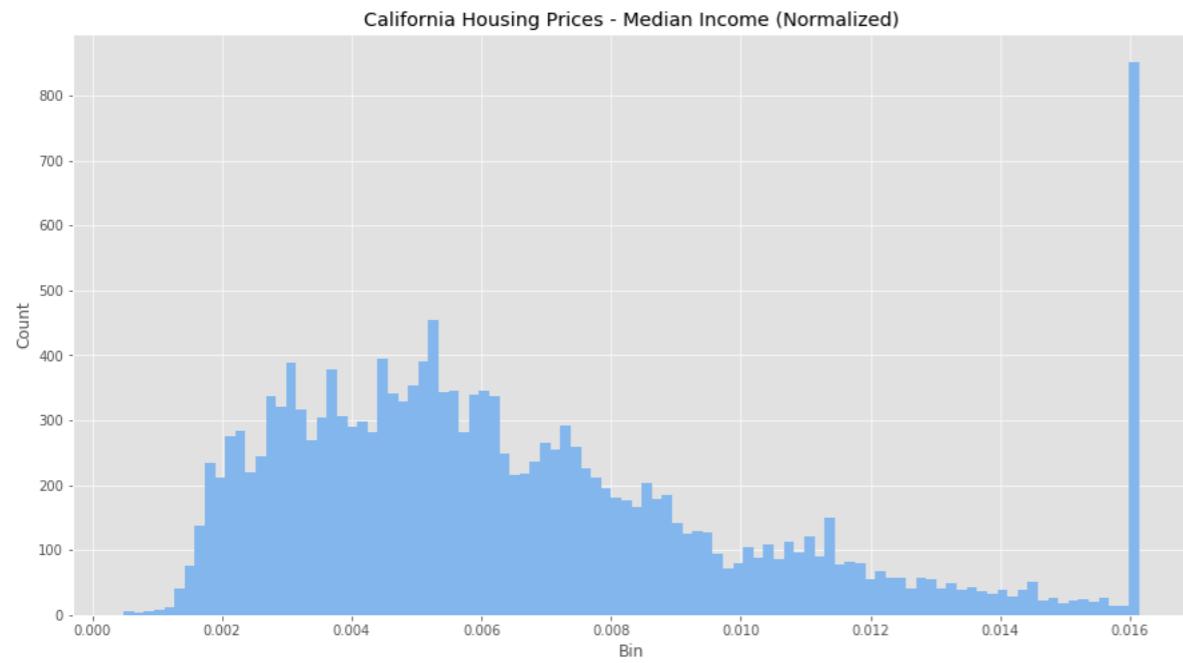
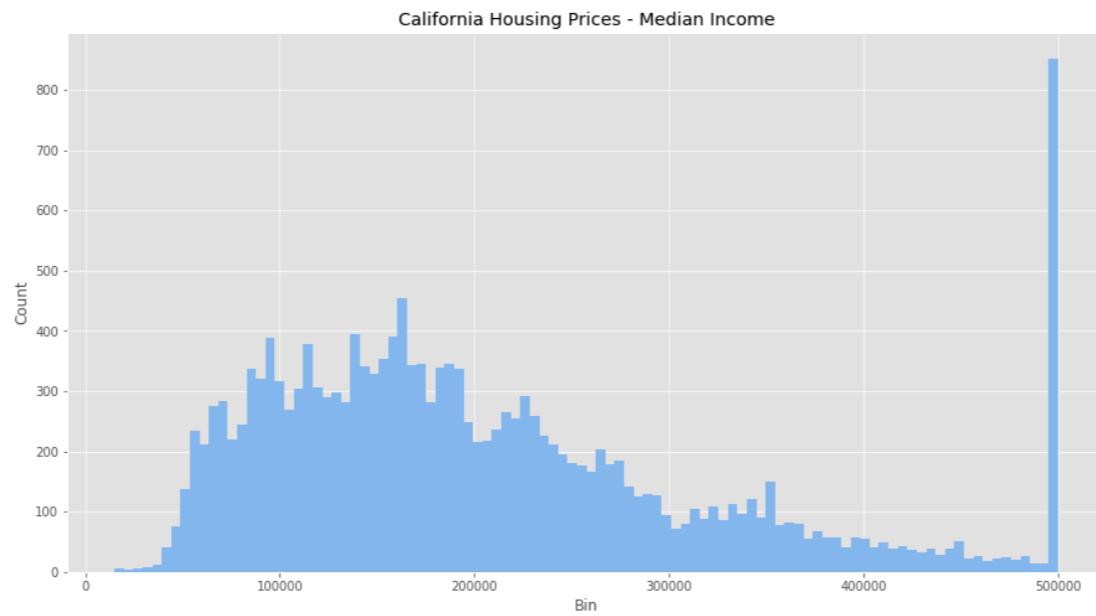
СТАНДАРТИЗАЦИЯ



СТАНДАРТИЗАЦИЯ



НОРМАЛИЗАЦИЯ VS СТАНДАРТИЗАЦИЯ



Категориальные переменные

Проблемы, которых никто не ждал [3]

ОБРАЗЕЦ ДАННЫХ

	age	job	marital	education	default
0	26	student	single	3	no
1	46	admin.	married	6	no
2	49	blue-collar	married	0	unknown
3	31	technician	married	6	no
4	42	housemaid	married	6	no

NUMERIC ENCODING

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign
0	26	8	2	3	0	0	0	1	4	1	901	1
1	46	0	1	6	0	2	0	0	1	3	208	2
2	49	1	1	0	1	2	2	1	4	3	131	5
3	31	9	1	6	0	0	0	0	3	3	404	1
4	42	3	1	6	0	2	0	1	7	1	85	1

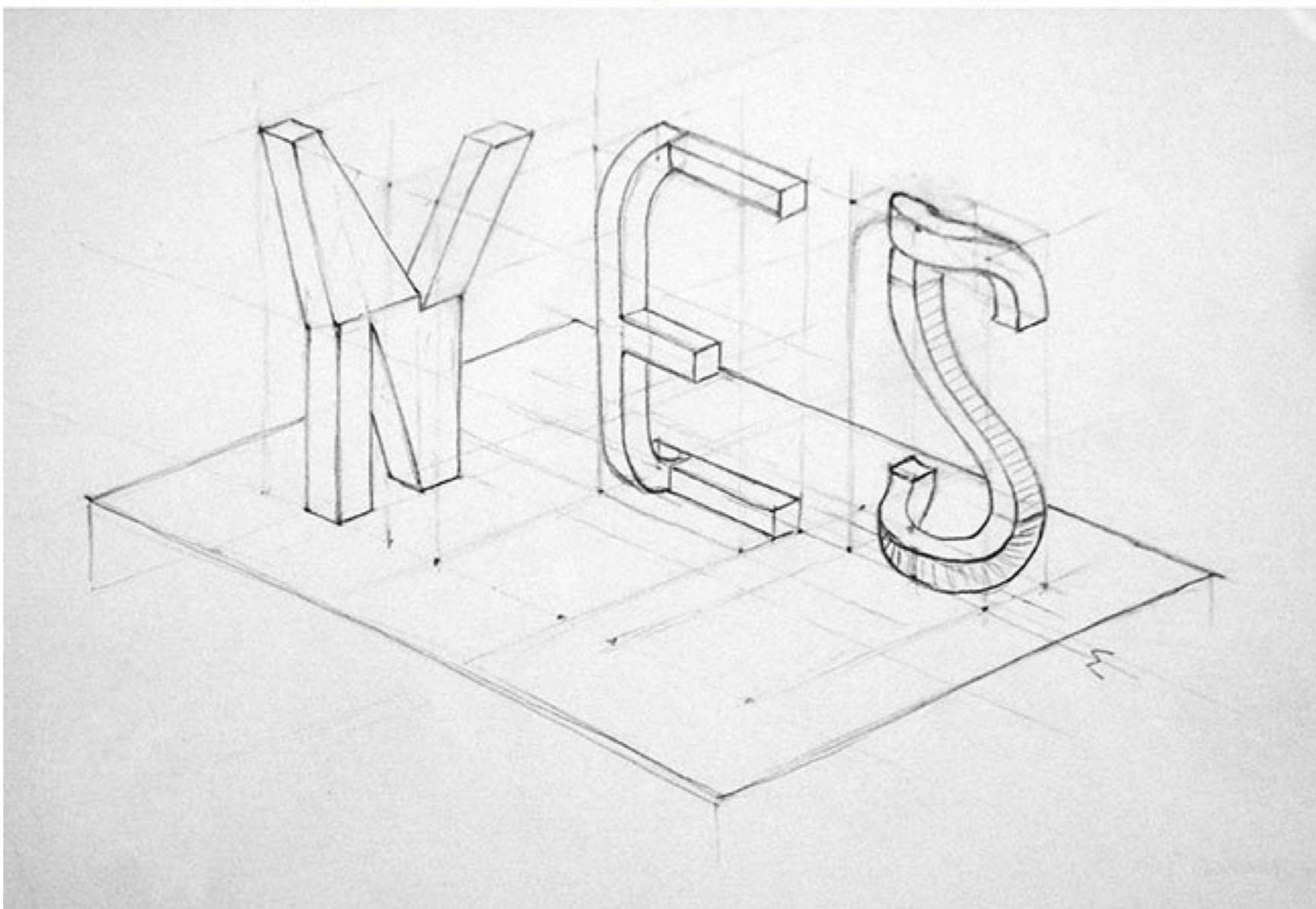
ONE HOT ENCODING

	0	1	2	3	4	5	6	7	8	9	...	43	44	45	46	47	48	49	50	51	52	
0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
1	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
2	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
4	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0

Количество переменных

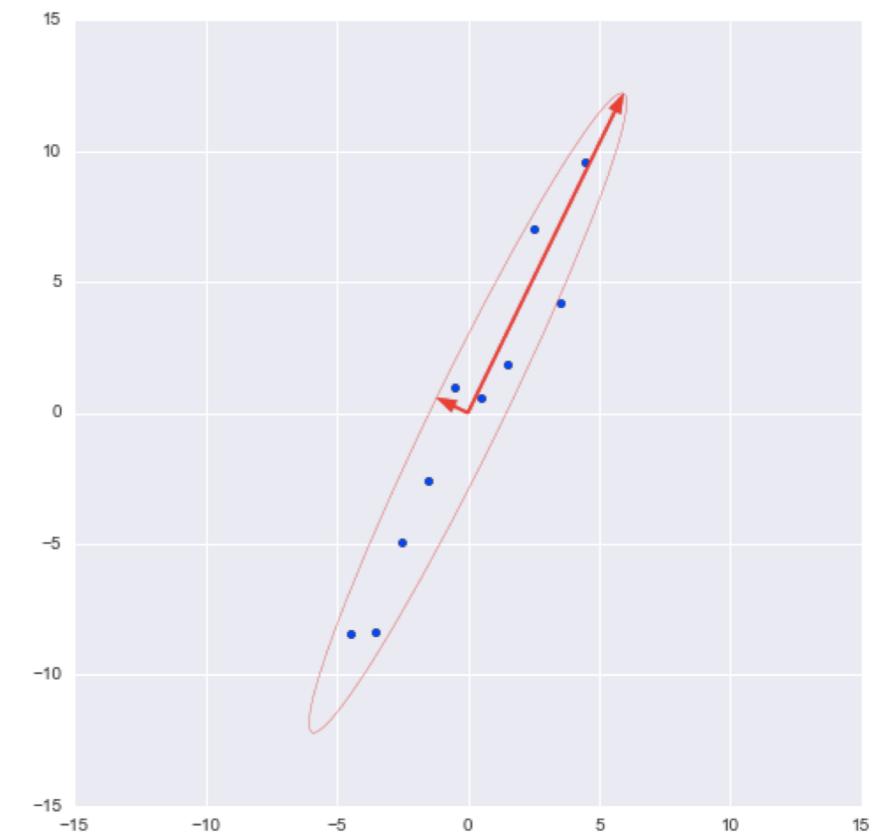
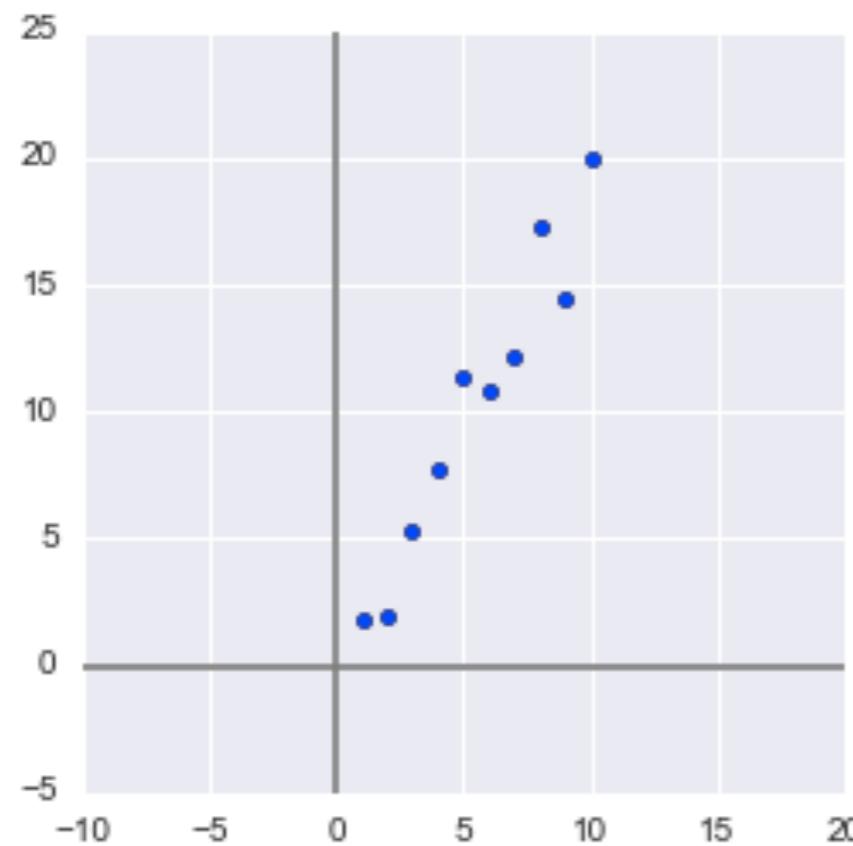
Проблемы, которых никто не ждал [4]

PRINCIPAL COMPONENT ANALYSIS



PCA

- Признаки могут объяснять друг друга. Причем неявно – корреляциями не поймать
- Для простоты - два признака, сильно коррелирующих друг с другом





НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ