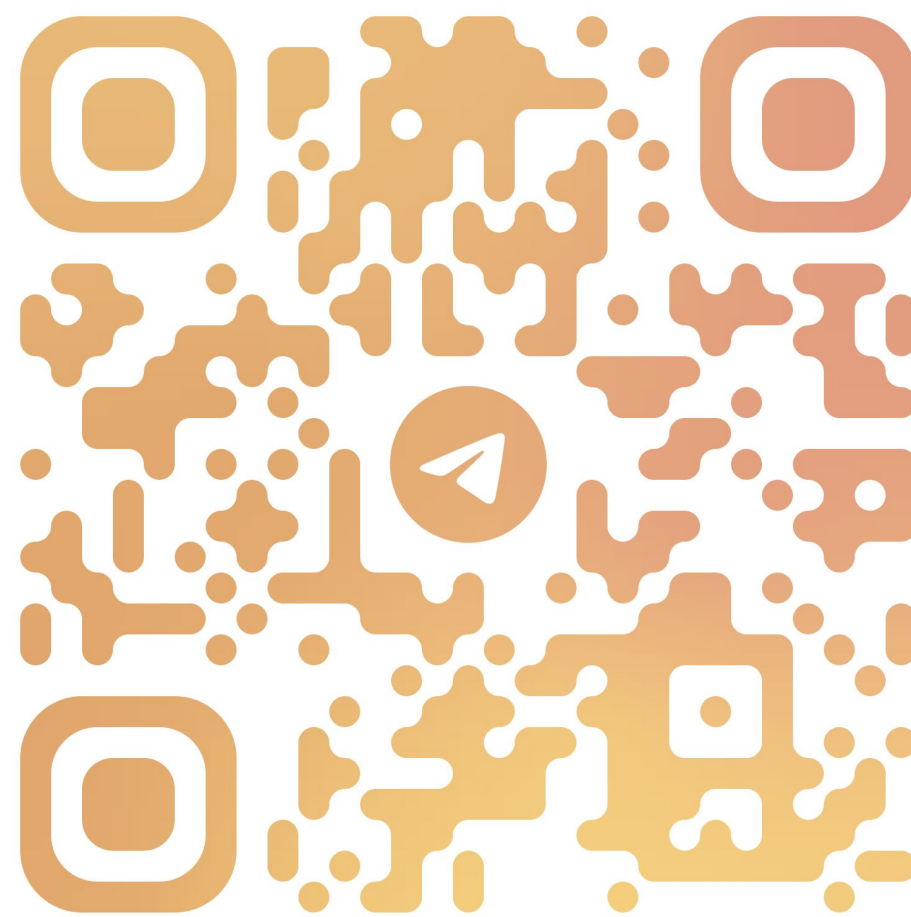


ML course.

Setup



Наш план

Пары:

лекции - каждую пятницу в 17.10, 1229

практики - раз в две недели очные семинары, в 18.50

защита лаб - после семинара/ онлайн в зуме

Наш план

Оценивание:

домашки - 60 баллов

рубежка - 10 баллов

семинары - 10 баллов

экзамен - 20 баллов

экзамен:

на 5 -- в виде проектного задания

на 4 и 3 -- обычный по билетам

Наш план

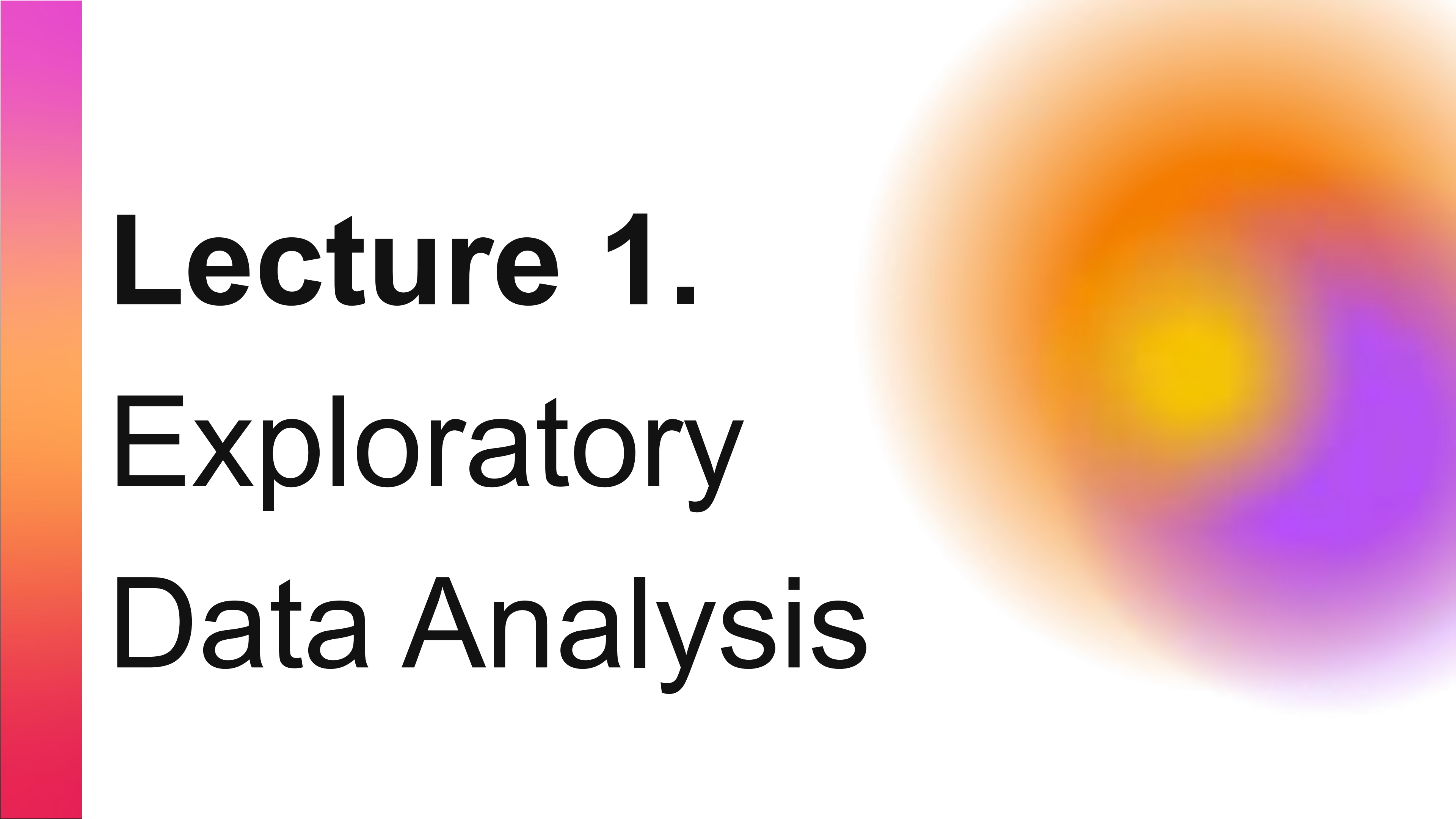
Домашек будет около 12 штук

Каждая стоит 100 баллов

Далее нормировка в 60 баллов

**Рубежка одна где-то в середине-конце ноября
(полупрактическая)**

Семинары: будет возможность решать задачи на семинарах и рассказывать их, показывать примеры визуализаций и кода



Lecture 1.

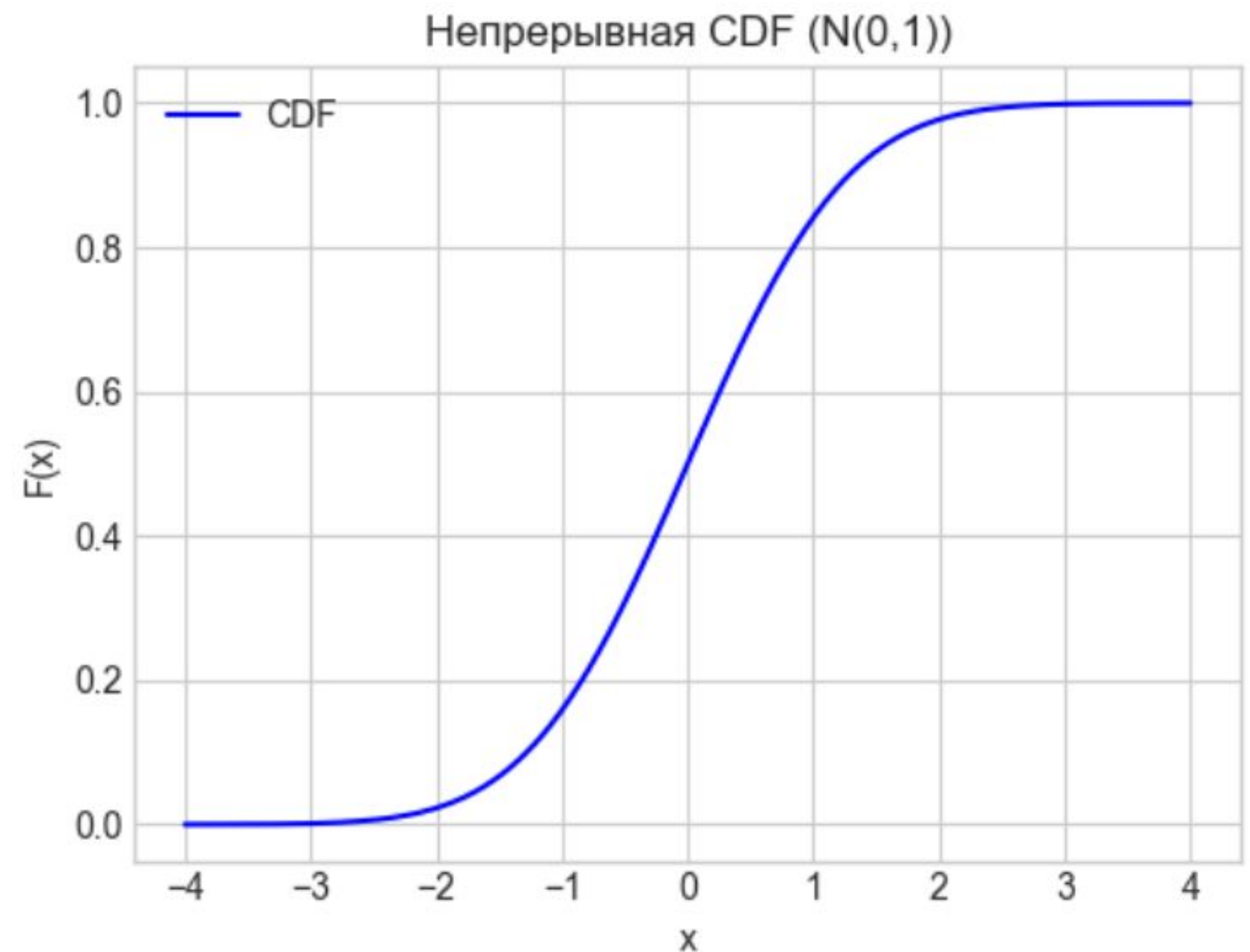
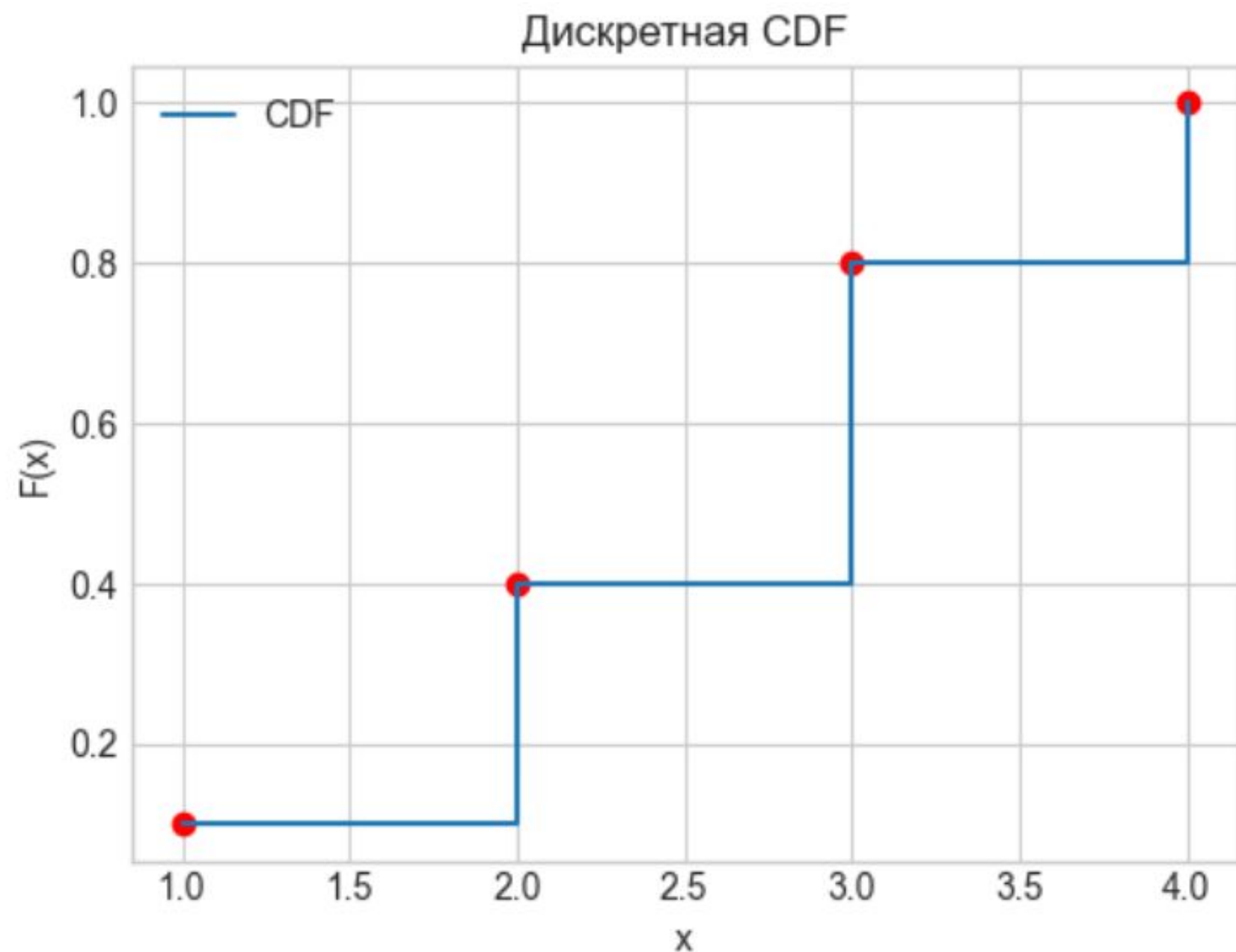
Exploratory

Data Analysis

Повторяем матстат.

Функция распределения.

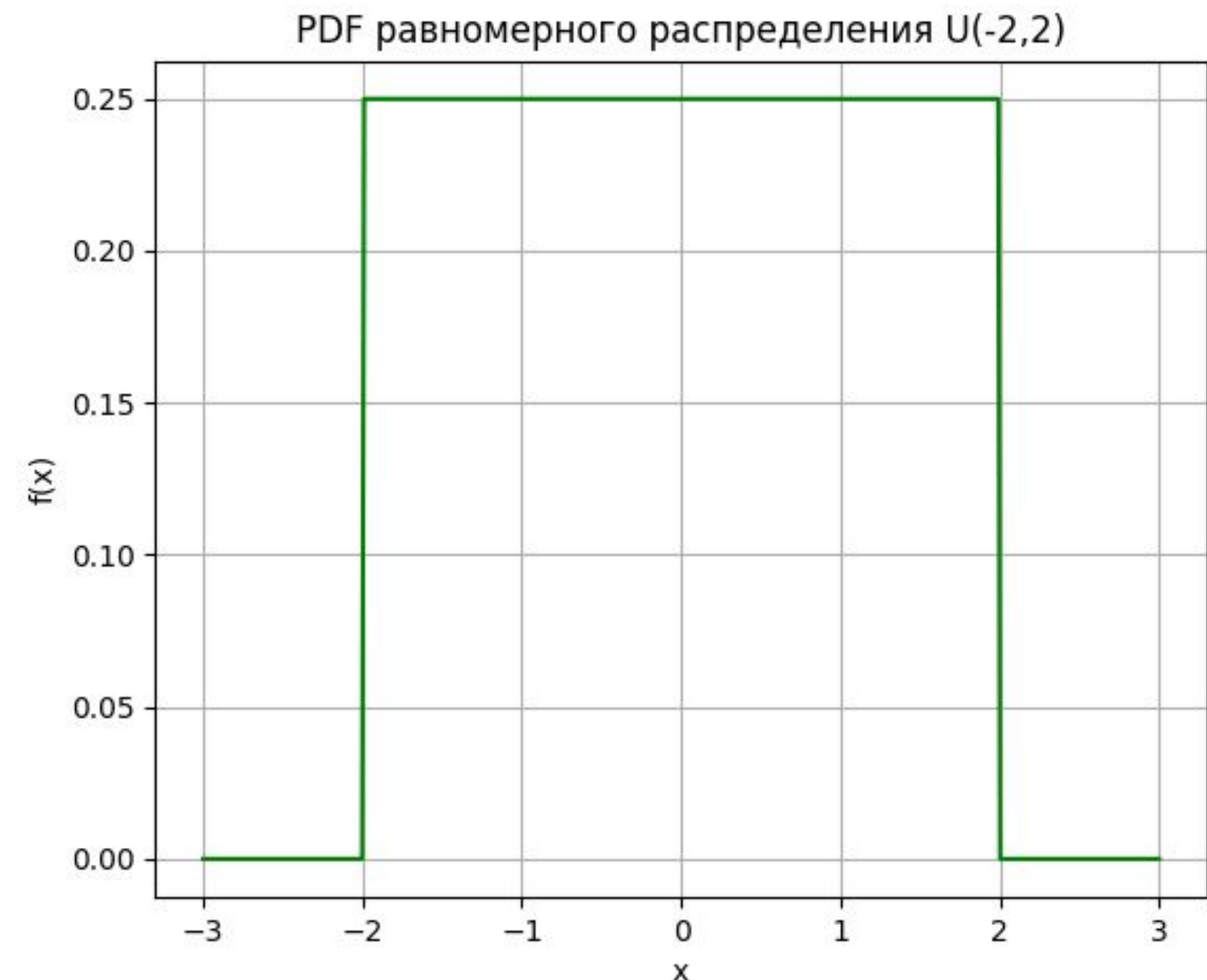
Функция распределения случайной величины ξ : $F_\xi(x) = P(\xi < x)$



Повторяем матстат.

Функция плотности.

Функция плотности случайной величины ξ : $f_{\xi}(x) = F'_{\xi}(x)$



Часто встречающиеся распределения

Биномиальное распределение.

Биномиальное распределение - дискретное распределение вероятностей случайной величины **X**, принимающей целочисленные значения **k=0,1,...,n** с вероятностями:

$$P(X = k) = C_n^k p^k (1 - p)^{n-k}$$

Обозначается **B(n,p)**, где:

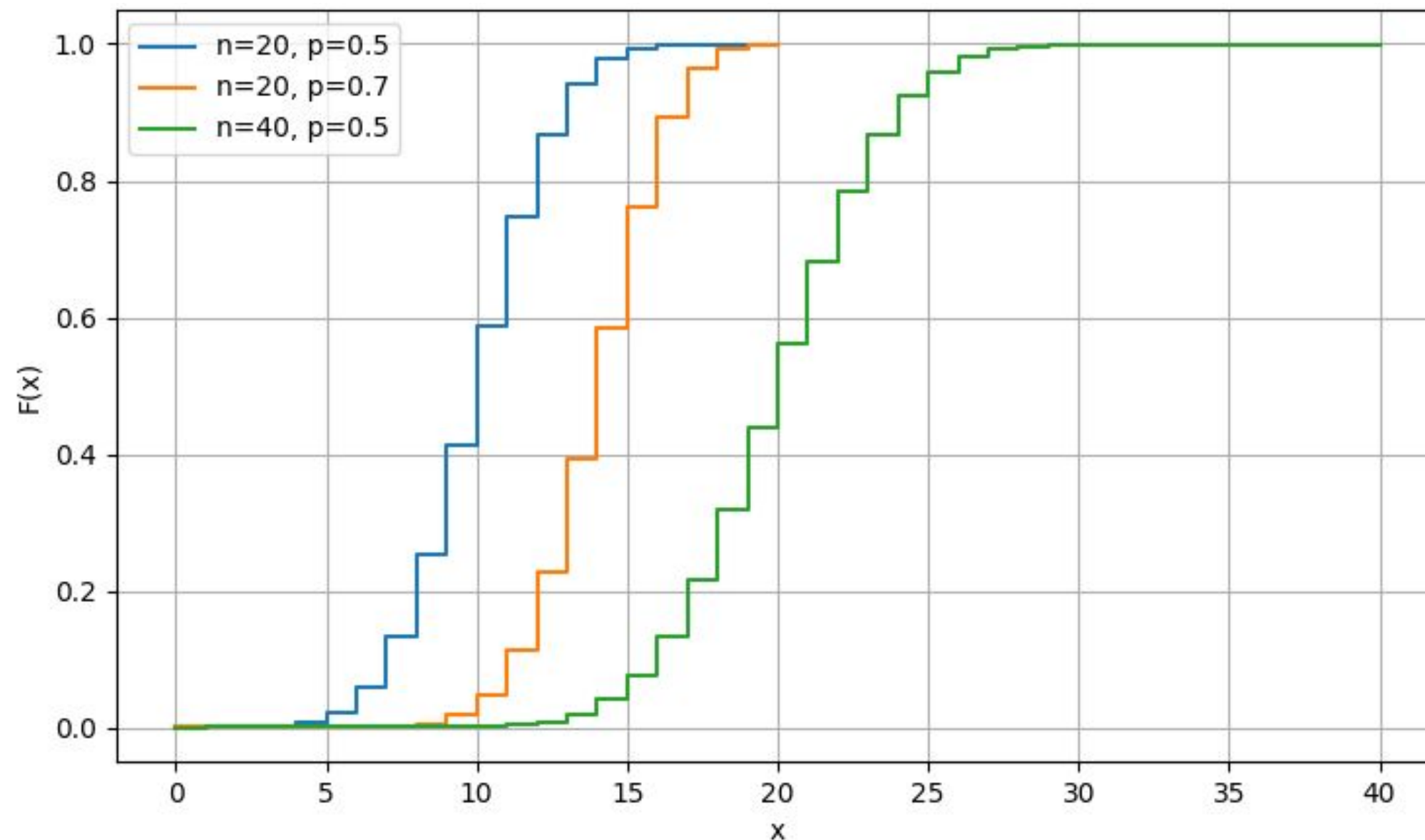
n - число испытаний

p - вероятность успеха

Часто встречающиеся распределения

Биномиальное распределение.

Функция распределения: $F(k) = \sum_{i=0}^{k-1} P(X = i) = \sum_{i=0}^{k-1} C_n^i p^i (1-p)^{n-i}$



Часто встречающиеся распределения

Распределение Пуассона.

Распределение Пуассона - дискретное распределение вероятностей случайной величины ξ , с параметром λ и функцией распределения:

$$F_{\xi}(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \lambda > 0$$

Обозначается $\Pi(\lambda)$.

Часто встречающиеся распределения

Распределение Пуассона.



Часто встречающиеся распределения

Равномерное распределение.

Случайная величина ξ имеет равномерное распределение $U(a,b)$ если ее плотность постоянна на $[a;b]$.

$$f_{\xi}(x) = \begin{cases} 0, & x < a \\ \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x > b \end{cases}$$

$$\text{При } x < a : F_{\xi}(x) = 0$$

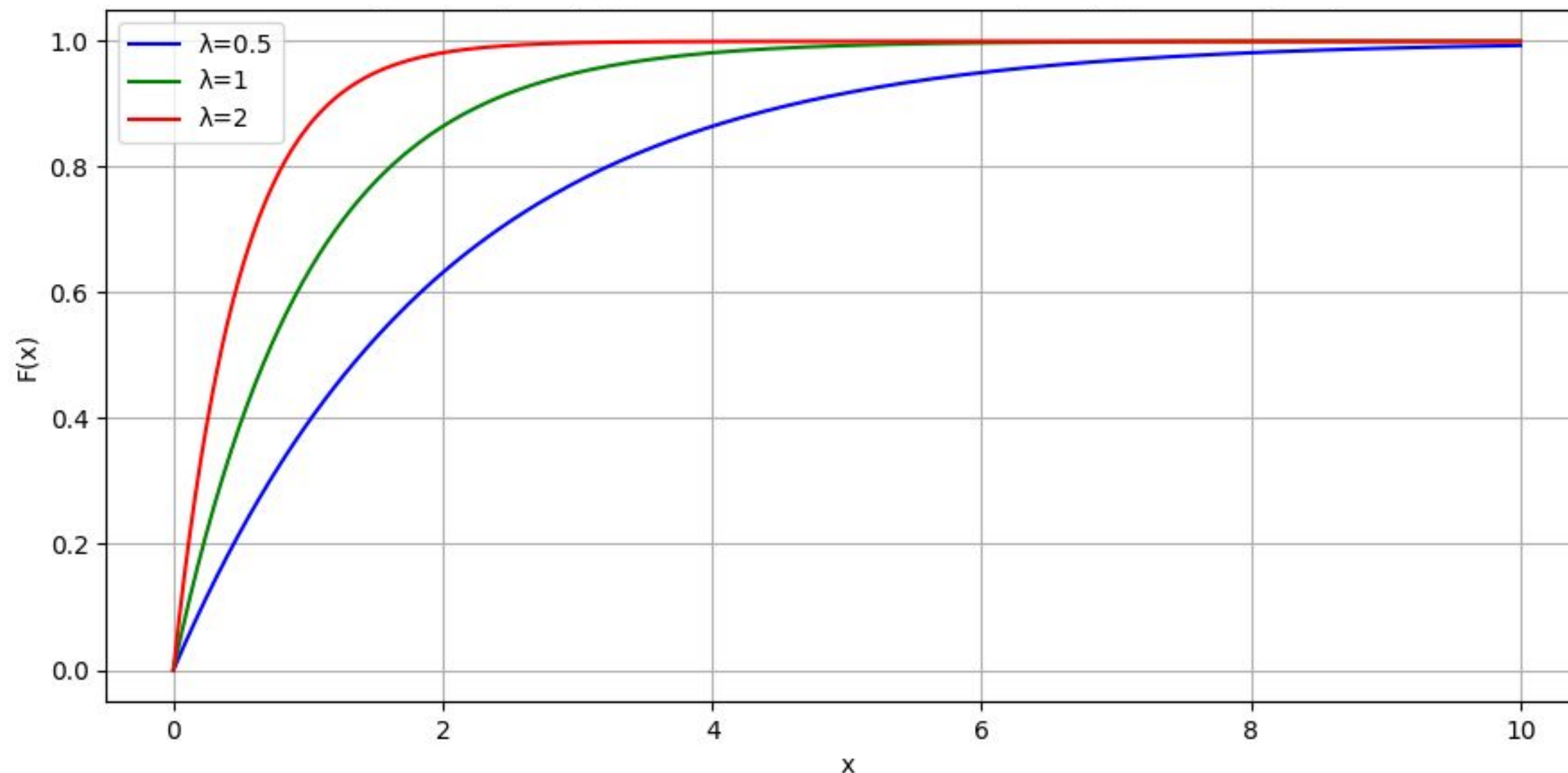
$$\text{При } a \leq x \leq b : F_{\xi}(x) = \int_a^x f_{\xi}(x) dx = \int_a^x \frac{1}{b-a} dx = \frac{x-a}{b-a}$$

$$\text{При } b < x : F_{\xi}(x) = 1$$

Часто встречающиеся распределения

Показательное распределение.

Случайная величина ξ имеет показательное распределение $E(\alpha)$, если ее плотность:

$$f_{\xi}(x) = \begin{cases} 0, & x < 0 \\ \alpha e^{-\alpha x}, & 0 \leq x \end{cases}$$


Часто встречающиеся распределения

Нормальное распределение.

Случайная величина ξ имеет нормальное распределение $N(a, \sigma^2)$

если ее плотность:

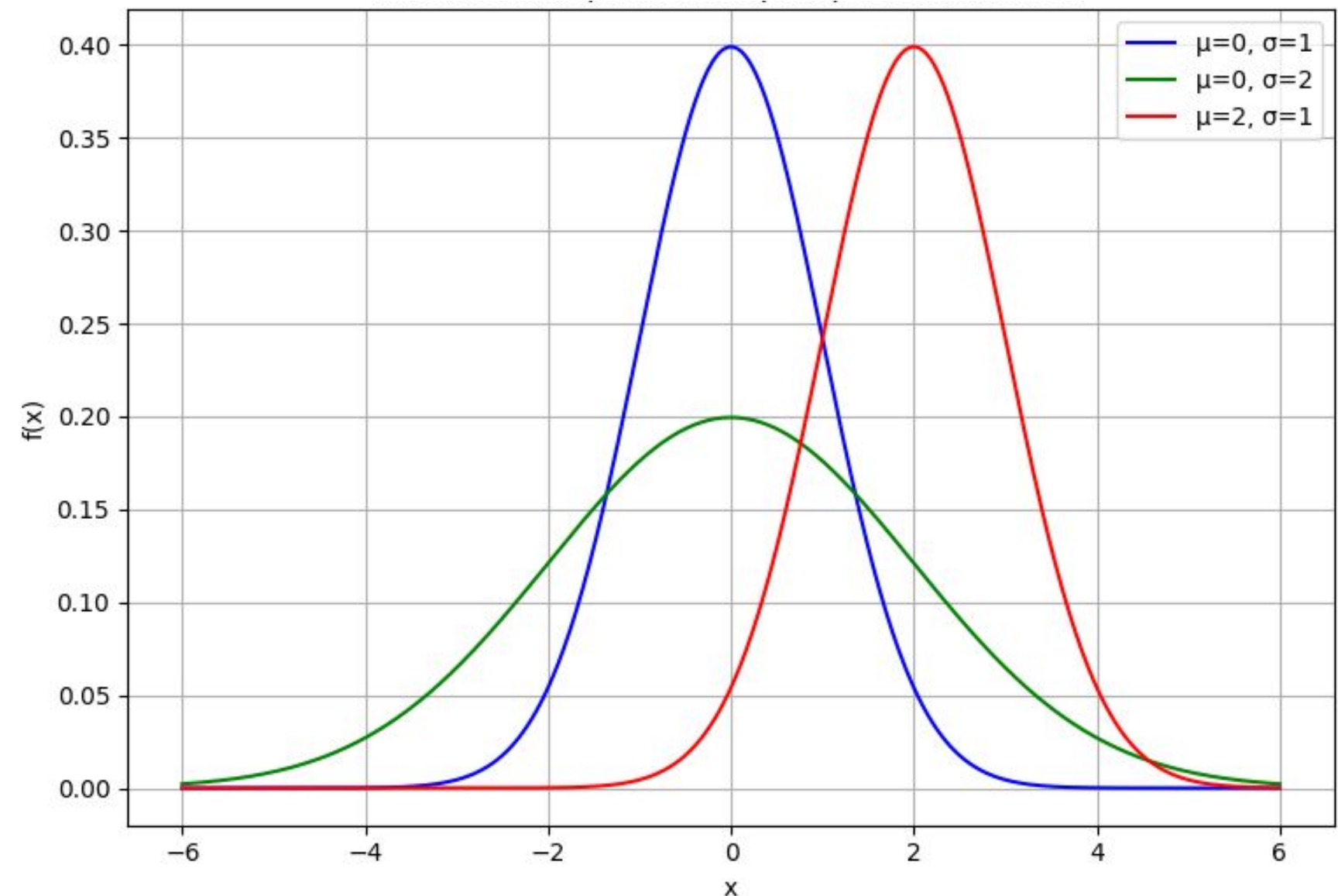
$$f_{\xi}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

$$F_{\xi}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-a)^2}{2\sigma^2}} dt$$

a - среднее

σ - среднеквадратичное

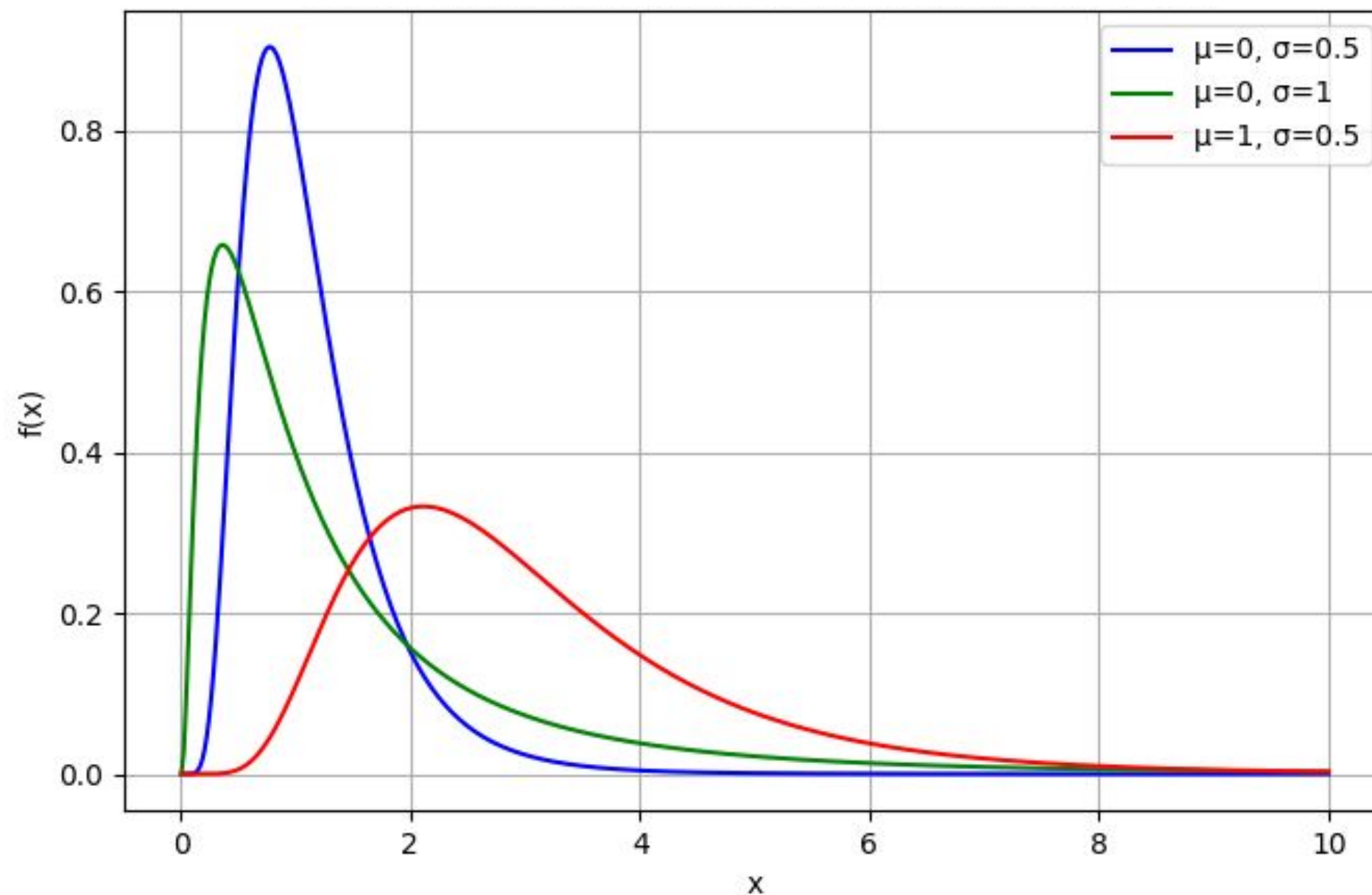
отклонение



Часто встречающиеся распределения

log-Нормальное распределение.

Если случайная величина X из нормально распределения, то $Y = e^X$ из **log-нормального** распределения.

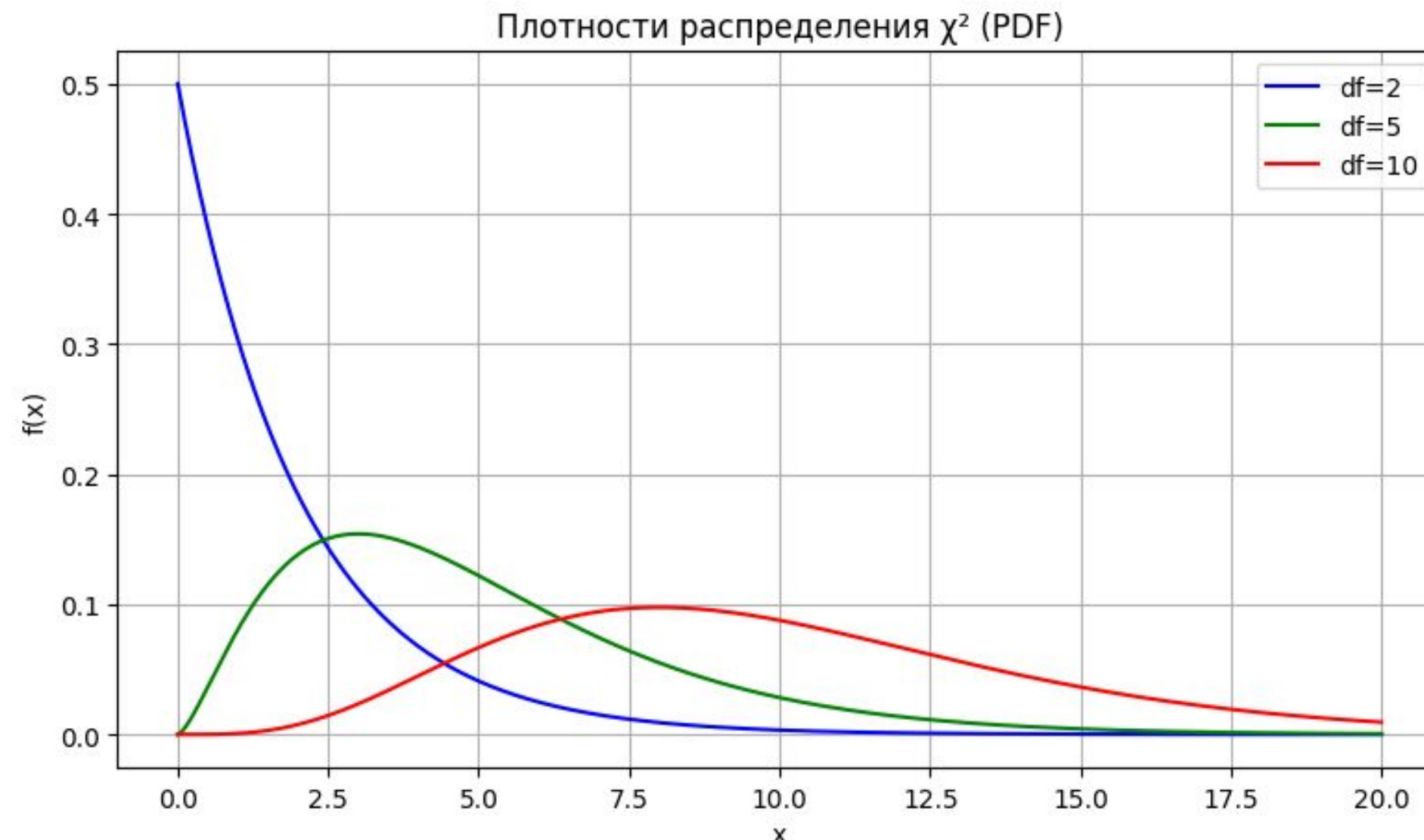


Часто встречающиеся распределения

Распределение "хи-квадрат"

Распределение "хи-квадрат" χ^2 с n степенями свободы n , называется распределением суммы квадратов независимых стандартных нормальных величин:

$$\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2 = \sum X_i^2$$



Часто встречающиеся распределения

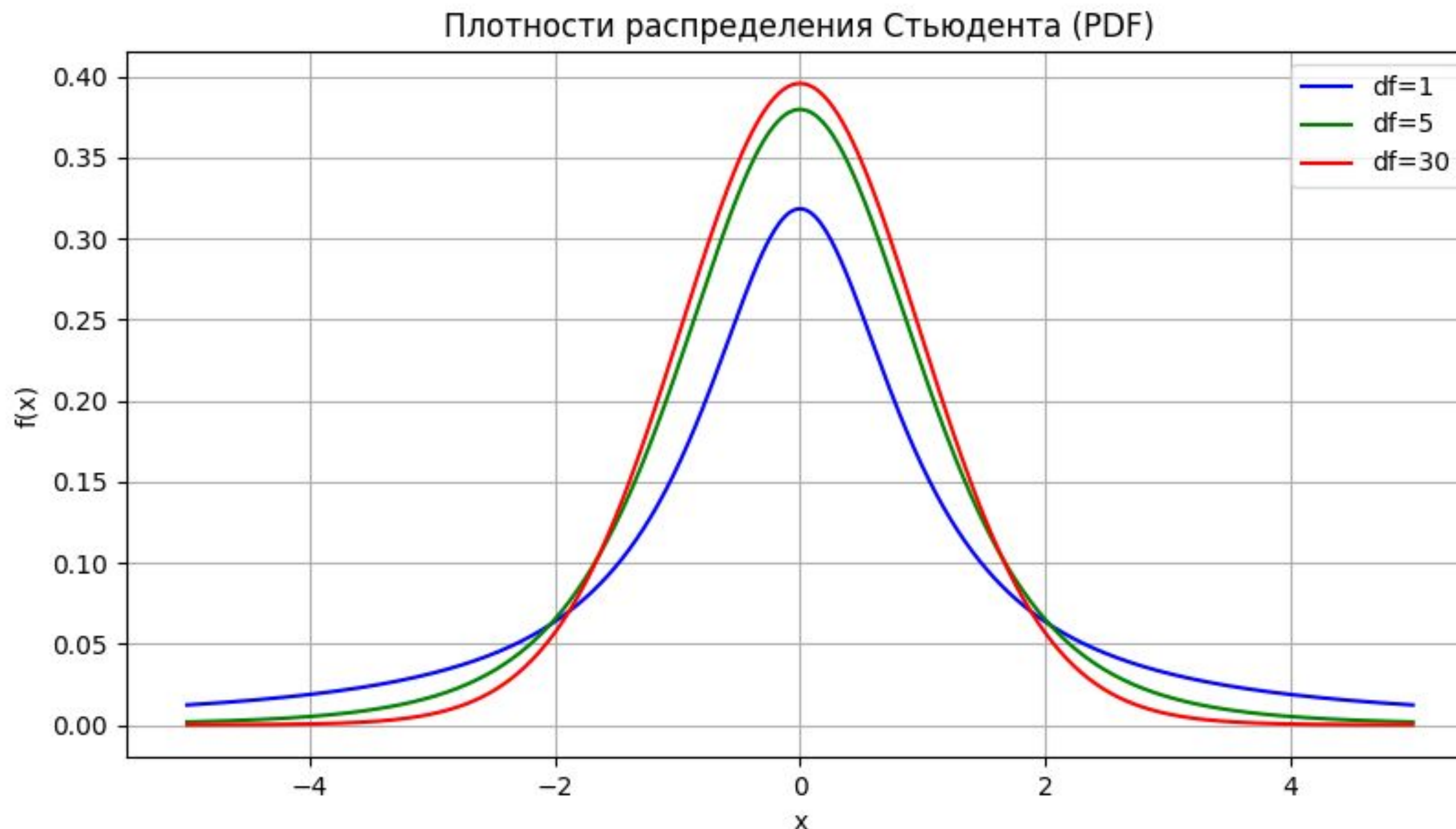
Распределение Стьюдента.

Распределением Стьюдента $T(k)$ с k степенями свободы называется распределение случайной величины:

$$t_k = \frac{X_0}{\sqrt{\frac{1}{k}(X_1^2 + \dots + X_k^2)}} = \frac{X_0}{\sqrt{\frac{\chi_k^2}{k}}}$$

Часто встречающиеся распределения

Распределение Стьюдента.

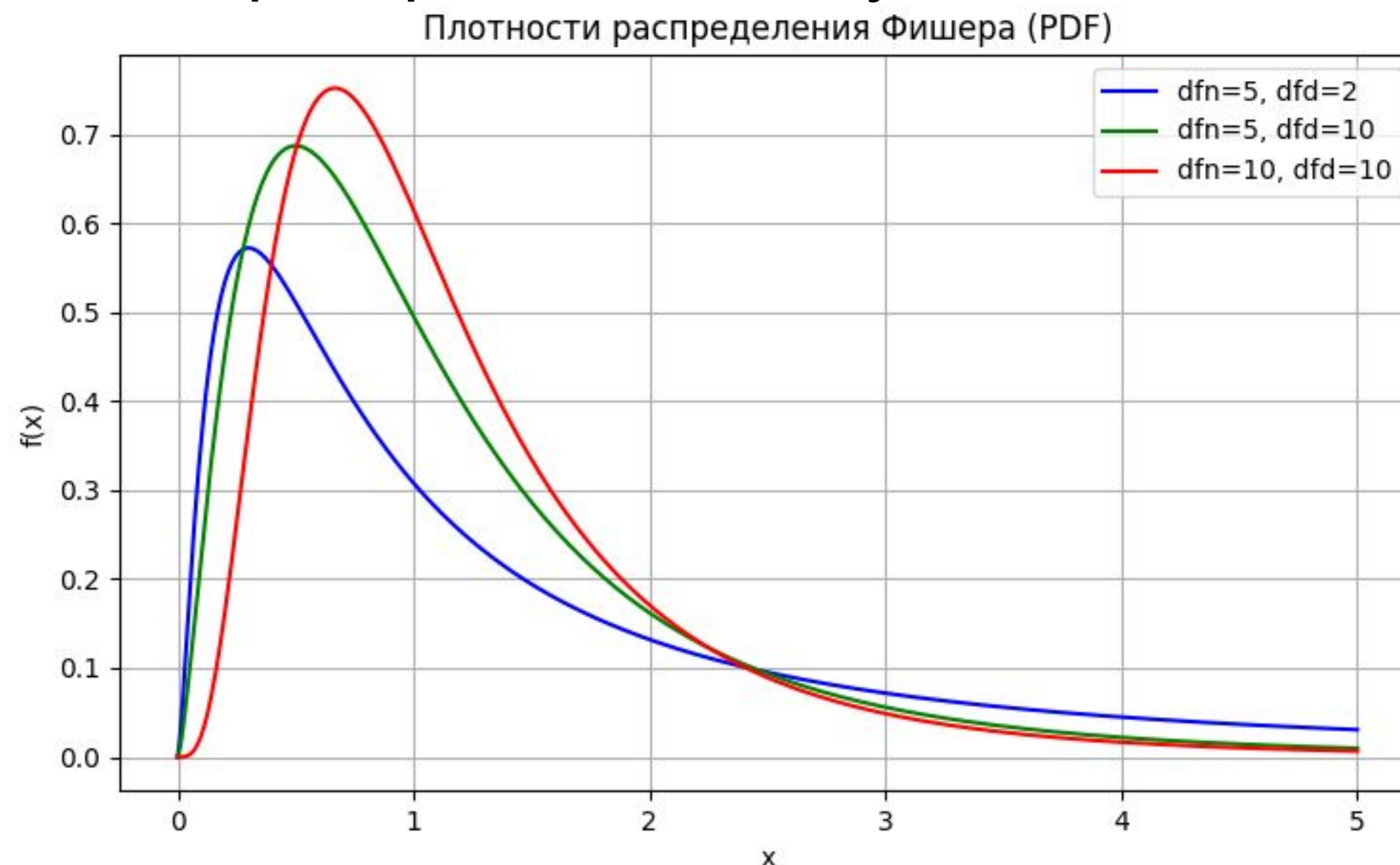


Часто встречающиеся распределения

Распределение Фишера.

Распределением **Фишера** $F(n,m)$ со степенями свободы n и m называется распределение случайной величины:

$$f_{n,m} = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}}$$



Характеристики распределений.

Математическое ожидание.

Дискретный случай:

Это взвешенное по вероятности среднее значение случайной величины.

Непрерывный случай:

Взвешивание будем проводить через функцию плотности распределения:

$$EX = \int_{-\infty}^{+\infty} x f_X(x) dx$$

Характеристики распределений.

Дисперсия, СКО.

Дисперсия - математическое ожидание квадрата отклонения случайной величины от её математического ожидания.

Дисперсия характеризует разброс случайной величины вокруг ее математического ожидания.

$$DX = DX^2 - (DX)^2$$

Среднеквадратическое отклонение:

$$\sigma X = \sqrt{DX}$$

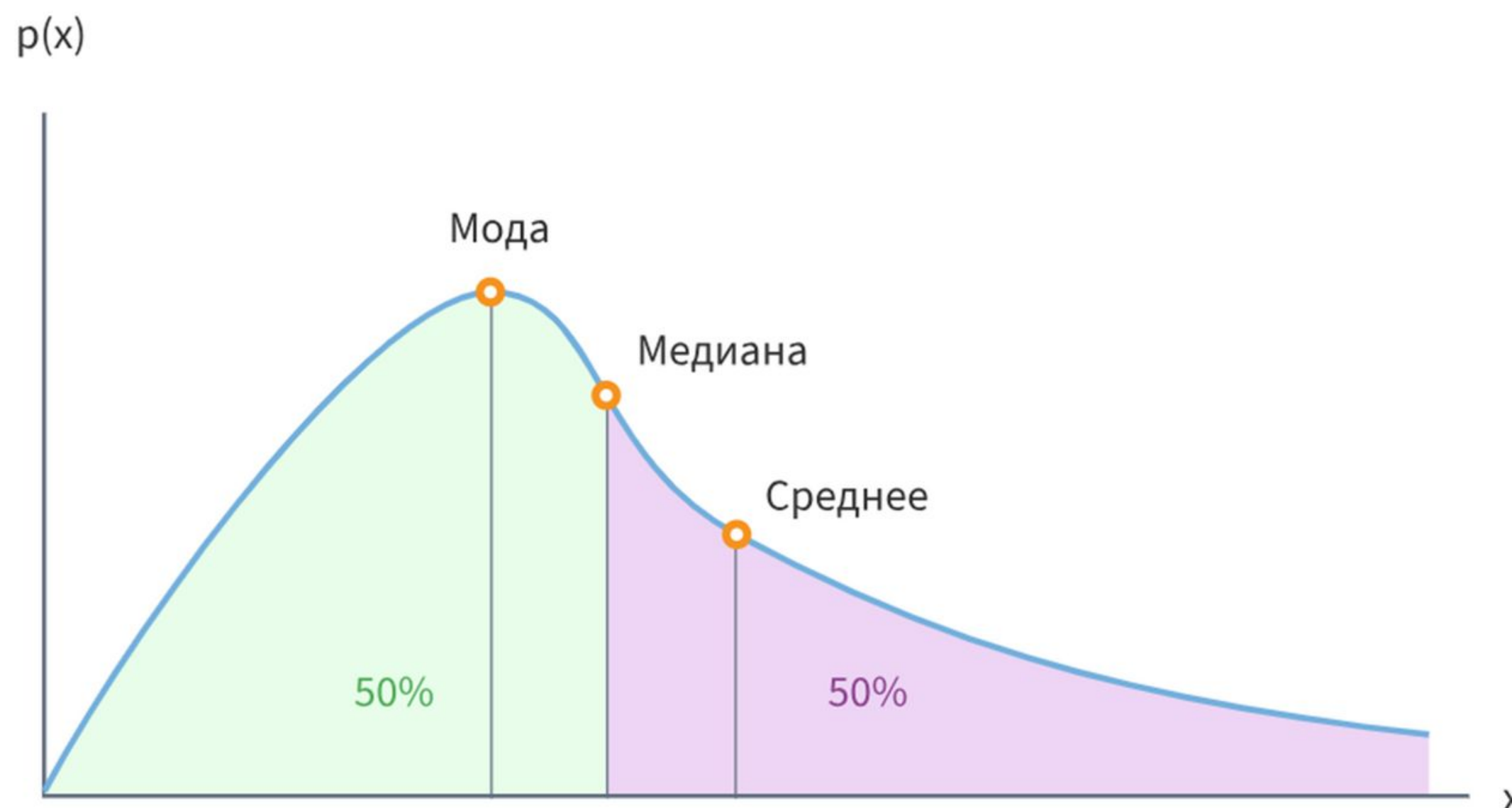
Характеристики распределений.

Мода, медиана, размах.

Медиана – значение, для которого значение функции распределения равно 0.5. То есть получения числа больше или меньше медианы равновероятно.

Мода – самое вероятное значение.

Размах – разность между максимальным и минимальным значением.

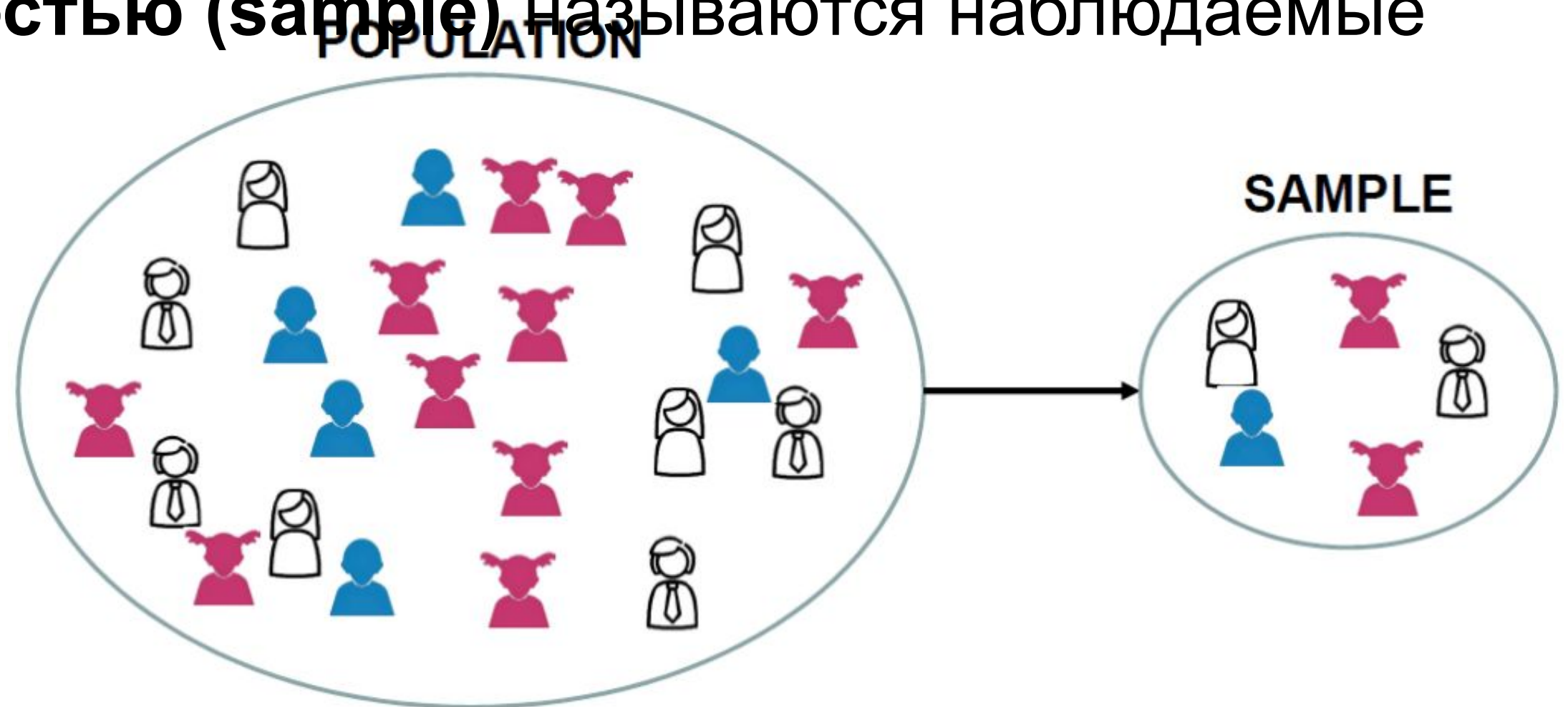


Выборки

Генеральная и выборочная совокупности.

Генеральной совокупностью (population) называются все результаты проведенных экспериментов.

Выборочной совокупностью (sample) называются наблюдаемые данные экспериментов.



Выборки

Репрезентативность.

Выборка называется **репрезентативной**, если её распределение совпадает с распределением генеральной совокупности.



Выборки

Репрезентативность. Пример.

Генеральная совокупность – вес вообще всех родившихся за последние три года щенков пуделя.

Выборка – опросили 10 заводчиков и собрали информацию о весе их щенков.

Является ли выборка репрезентативной?

Выборки

Характеристики выборок.

Выборочным средним называется

величина: $\bar{X} = \frac{1}{n} \sum X_i$

Выборочной дисперсией называется

величина: $D^* = \frac{1}{n} \sum (X_i - \bar{X})^2 = \frac{1}{n} \sum X_i^2 - \bar{X}^2$

Исправленной выборочной дисперсией называется

величина: $S^2 = \frac{n}{n-1} D^* = \frac{1}{n-1} \sum (X_i - \bar{X})^2$

Что делать с датасетом?

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750	male
1	Adelie	Torgersen	39.5	17.4	186.0	3800	female
2	Adelie	Torgersen	40.3	18.0	195.0	3250	female
4	Adelie	Torgersen	36.7	19.3	193.0	3450	female
5	Adelie	Torgersen	39.3	20.6	190.0	3650	male
6	Adelie	Torgersen	38.9	17.8	181.0	3625	female
7	Adelie	Torgersen	39.2	19.6	195.0	4675	male
8	Adelie	Torgersen	34.1	18.1	193.0	3475	Unknown
9	Adelie	Torgersen	42.0	20.2	190.0	4250	Unknown
10	Adelie	Torgersen	37.8	17.1	186.0	3300	Unknown
11	Adelie	Torgersen	37.8	17.3	180.0	3700	Unknown

Главные шаги препроцессинга

Анализ присутствующих фичей.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 344 entries, 0 to 343
```

```
Data columns (total 7 columns):
```

#	Column	Non-Null Count	Dtype
---	-----	-----	-----
0	species	344 non-null	category
1	island	344 non-null	category
2	bill_length_mm	342 non-null	float64
3	bill_depth_mm	342 non-null	float64
4	flipper_length_mm	342 non-null	float64
5	body_mass_g	342 non-null	float64
6	sex	333 non-null	category

```
dtypes: category(3), float64(4)
```

```
memory usage: 12.3 KB
```

Переменные могут
быть:

- Числовыми
- Номинальными
- Порядковыми

Главные шаги препроцессинга

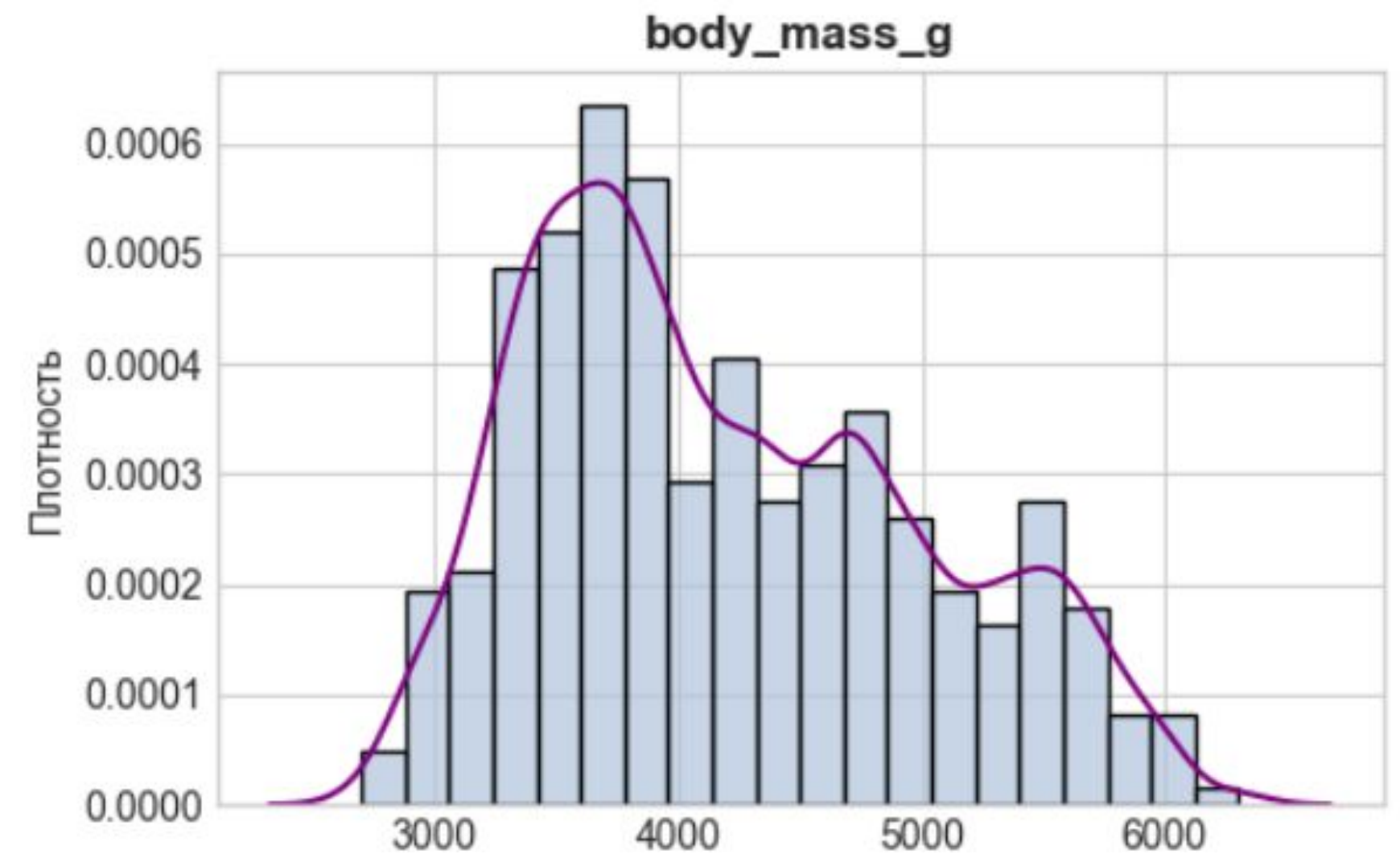
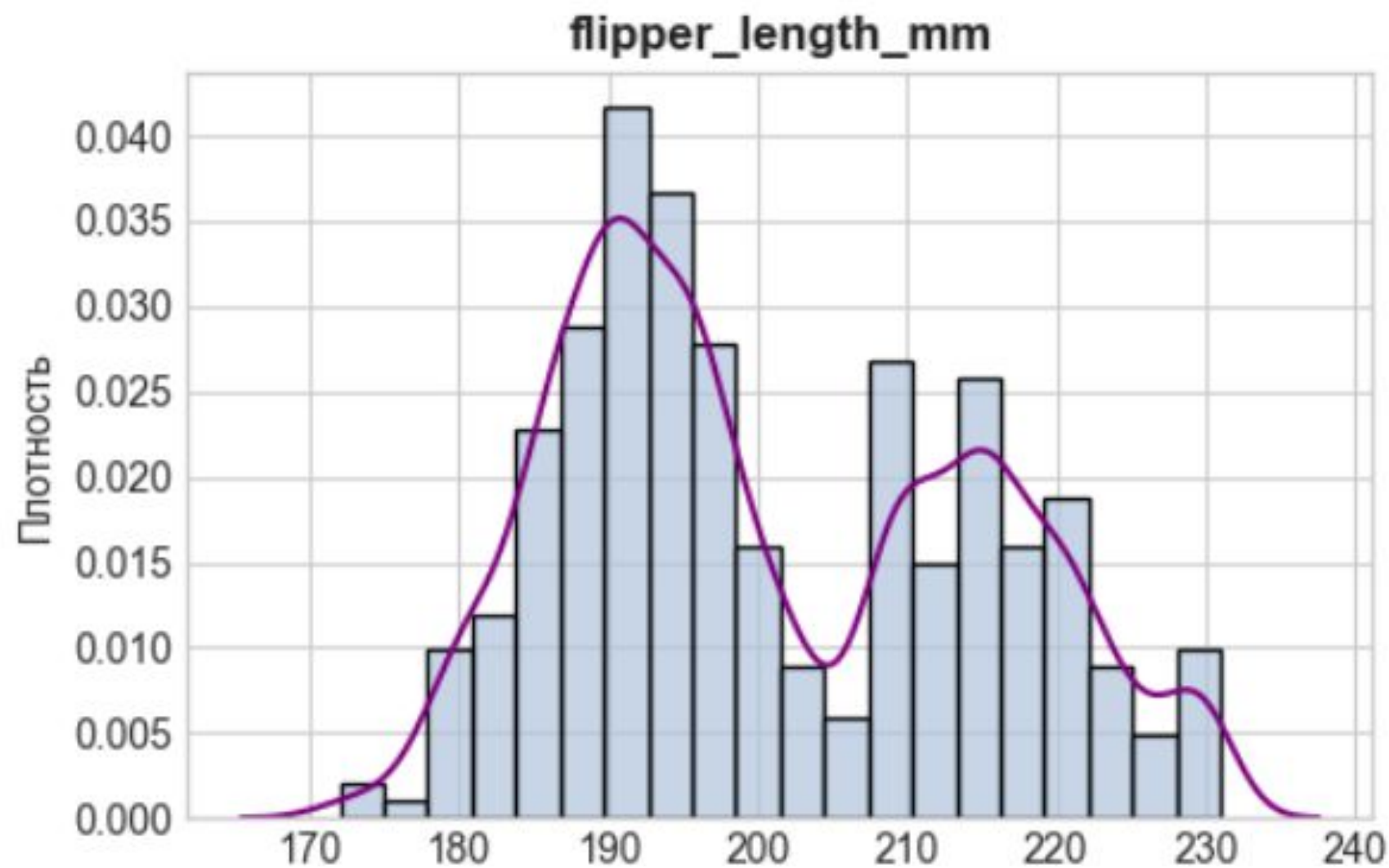
Анализ присутствующих фичей.

К какому типу относятся фичи?

ID	Возраст	Рост	Пол	Образование	Работает	Доход_тыс
1	25	175.5	М	Бакалавр	1	50
2	30	180.2	Ж	Магистр	1	80
3	22	168.0	Ж	Бакалавр	0	30
4	35	172.5	М	Доктор	1	100
5	28	177.0	М	Магистр	0	60

Главные шаги препроцессинга

Анализ распределений переменных.



Главные шаги препроцессинга

Анализ незаполненных значений.

Стратегии работы с NaN:

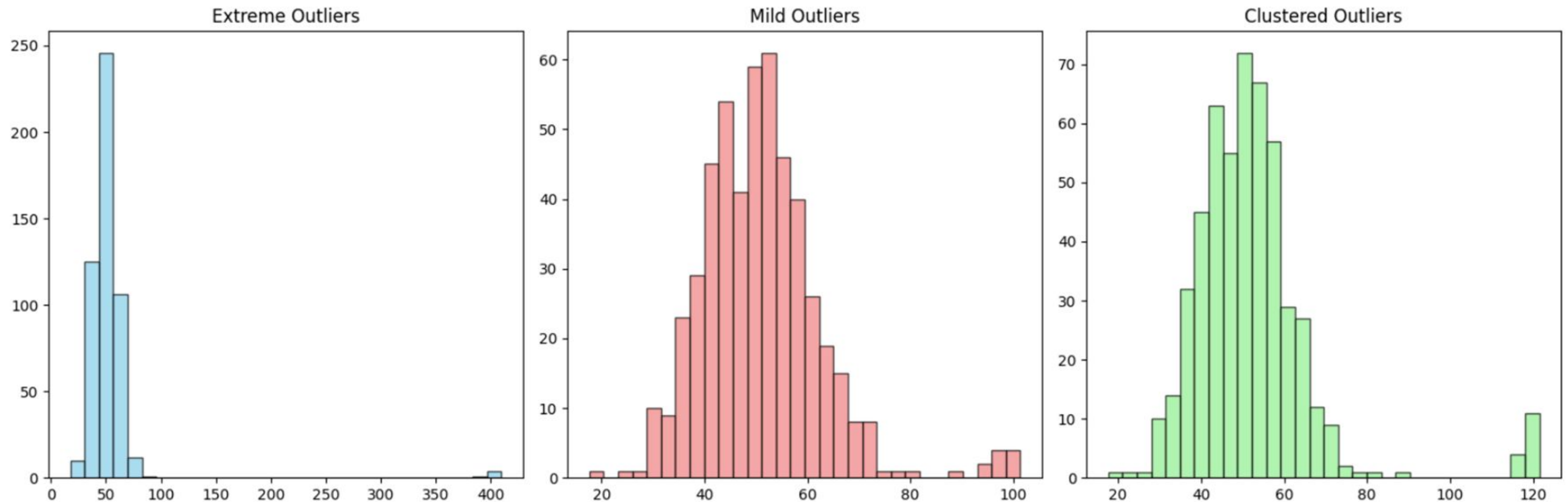
1. Удалить все строки, в которых есть NaN.
2. Заполнить Средним/Медианой/Модой.
3. Интерполяция.
4. Заполнение на основе соседних данных.

Всегда ли мы можем удалить/заполнить данные?

Главные шаги препроцессинга

Анализ аутлаеров.

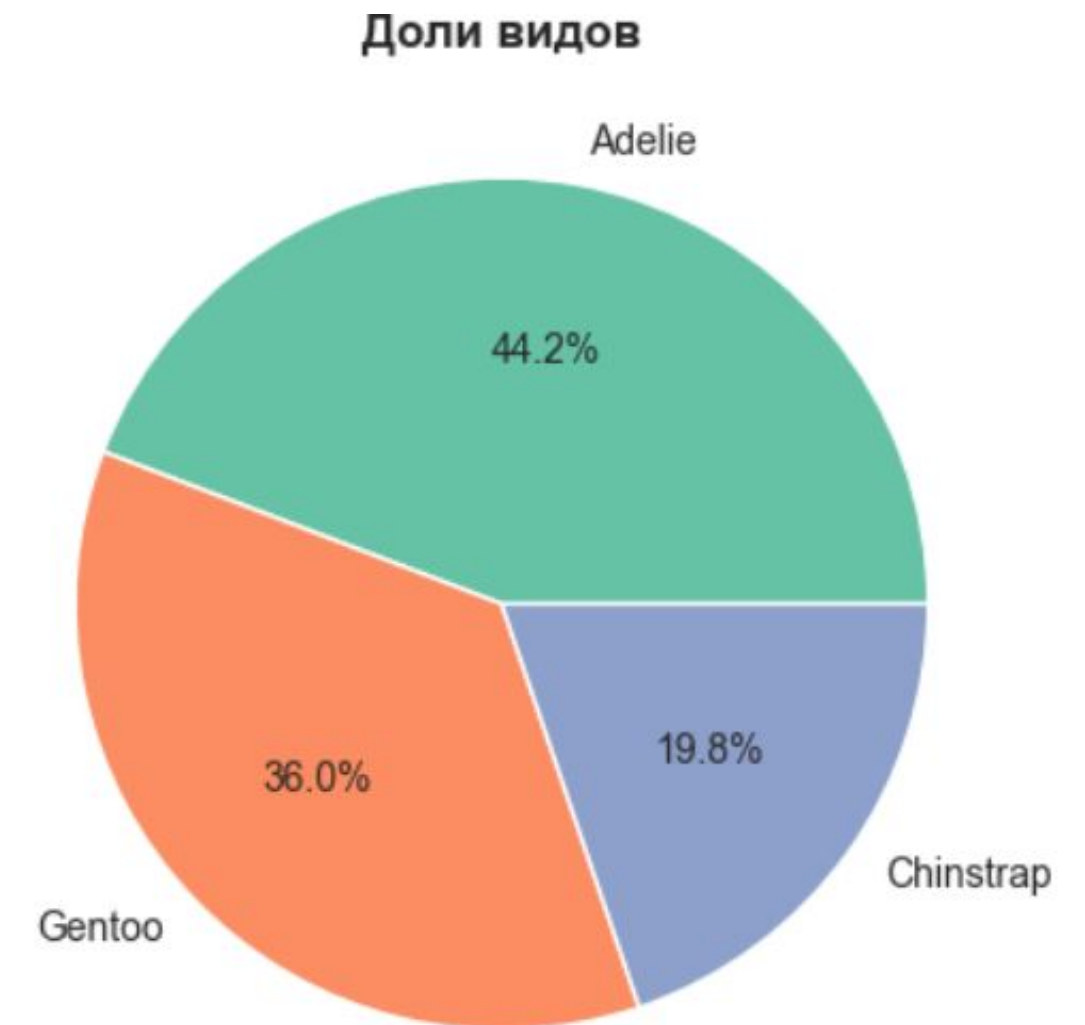
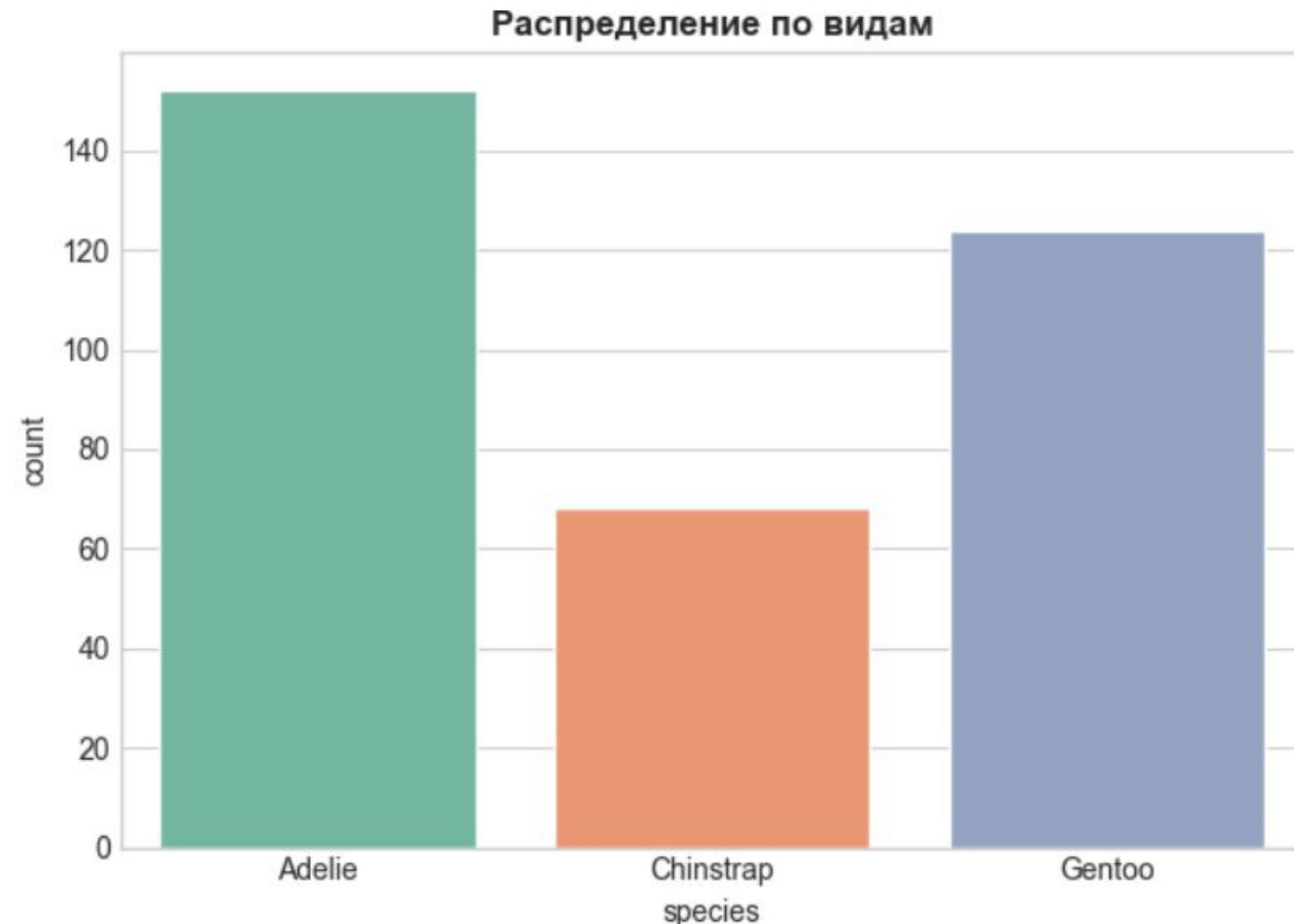
Аутлаеры (выбросы) - группа значений, выделяющаяся из общей выборки.



Главные шаги препроцессинга

Анализ категориальных переменных.

Категориальные переменные - принимают только определенный набор значений.
Значения не сравнимы между собой.



Главные шаги препроцессинга

Анализ категориальных переменных.

Методы машинного обучения работают с **числовыми значениями**.

Как сделать из категориальной переменной числовую?

color	color_code
red	0
green	1
blue	2

Почему это плохо?

Главные шаги препроцессинга

Анализ категориальных переменных.

One hot encoding - это метод представления категориальных данных в виде бинарных векторов, где каждая уникальная категория кодируется отдельным разрядом: значение категории обозначается 1 в соответствующем разряде, а все остальные разряды принимают значение 0.

color	is_blue	is_green	is_red
red	0	0	1
green	0	1	0
blue	1	0	0

Главные шаги препроцессинга

Анализ порядковых переменных.

Порядковые переменные принимают значения с естественным порядком. При этом расстояния между значениями часто неизвестны и не всегда равны.

ID	Name	Education Level	Income Range
1	John Smith	High School	\$20k-\$40k
2	Alice Brown	Bachelor's	\$40k-\$60k
3	Bob White	Master's	\$60k-\$80k
4	Carol Green	PhD	\$80k-\$100k
5	Dave Blue	Bachelor's	\$40k-\$60k
6	Emma Black	High School	\$20k-\$40k

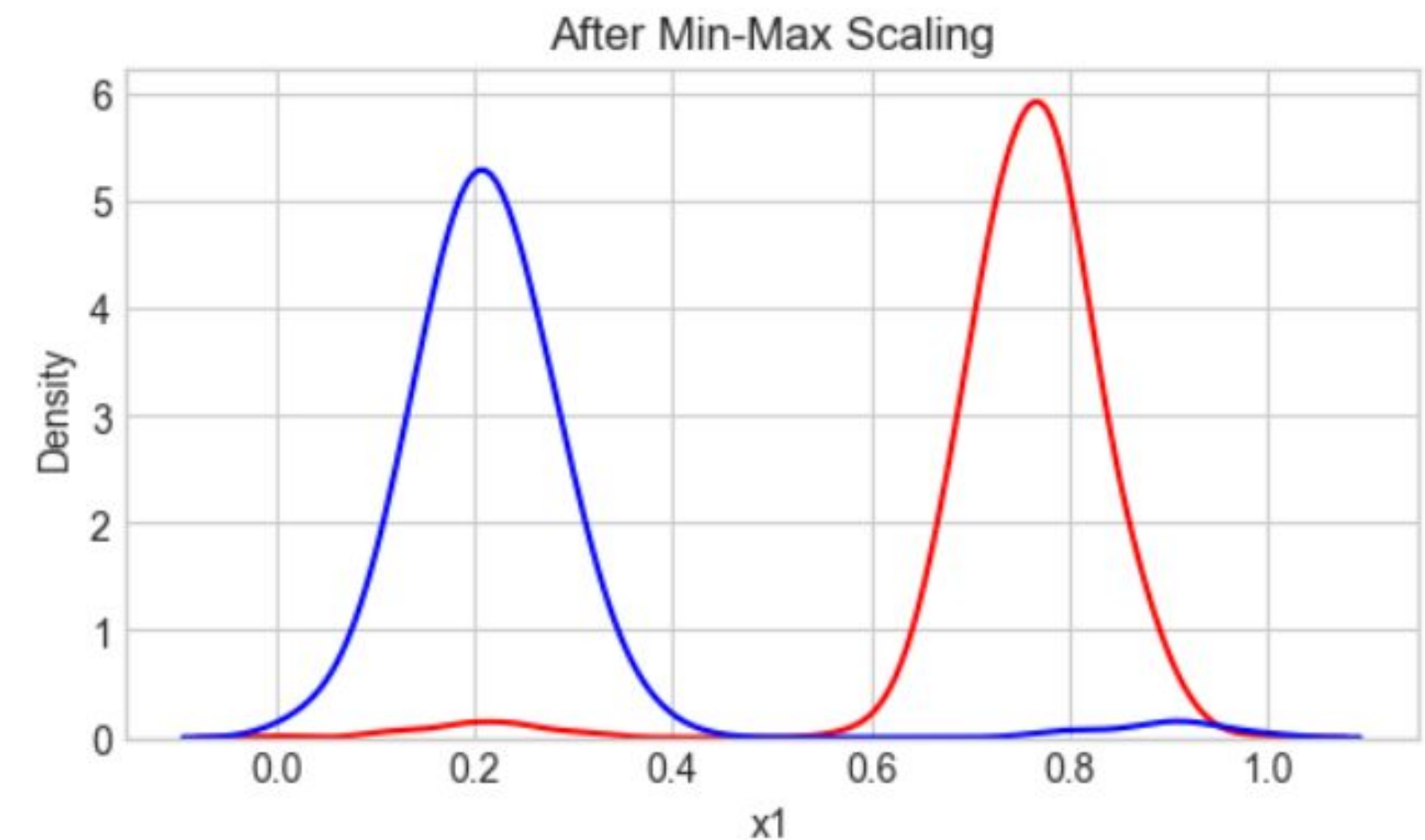
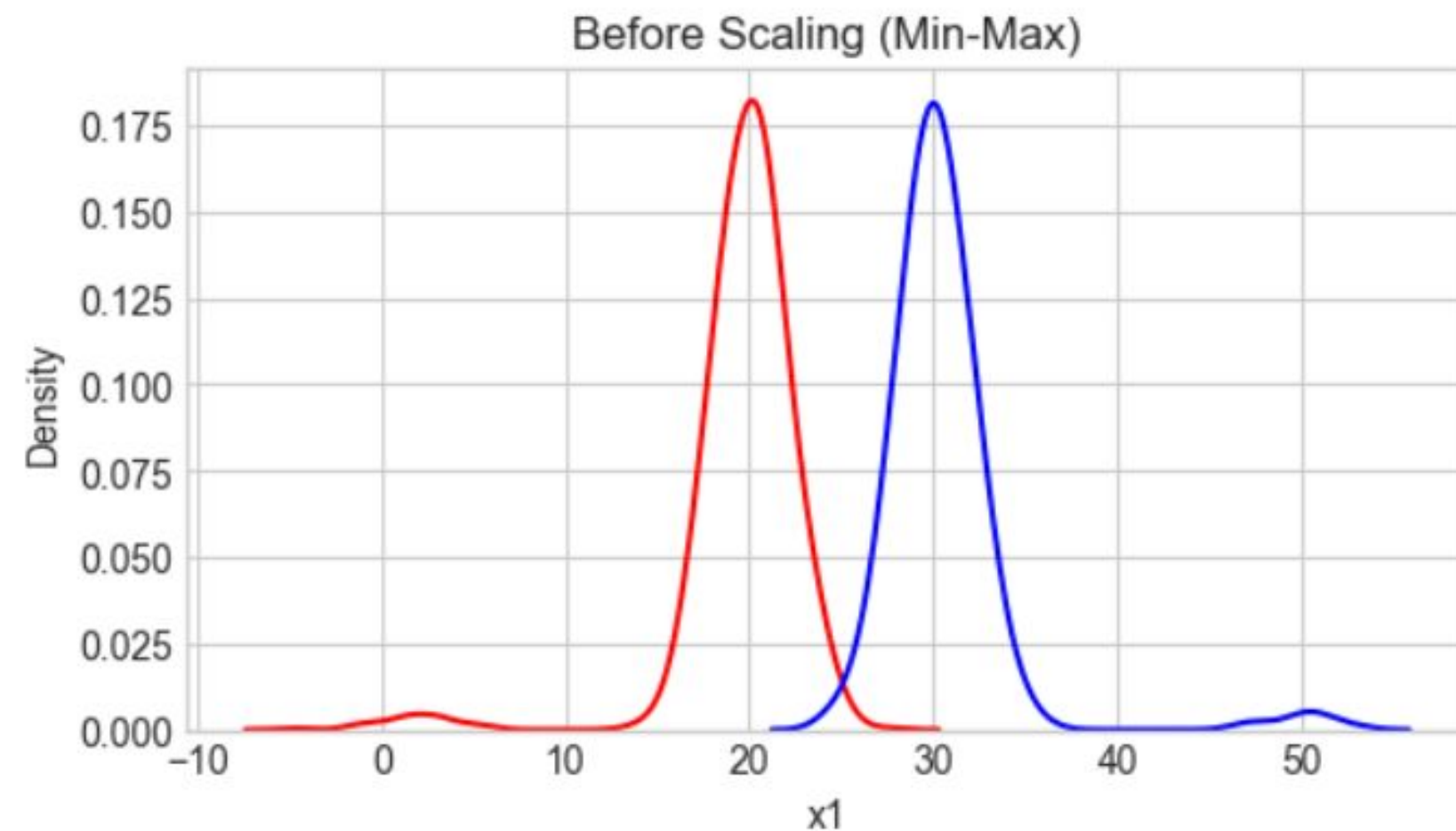
Главные шаги препроцессинга

Нормализация данных.

Модели машинного обучения и стат модели хотят, чтобы все фичи были равно важны при учёте.

Главные шаги препроцессинга

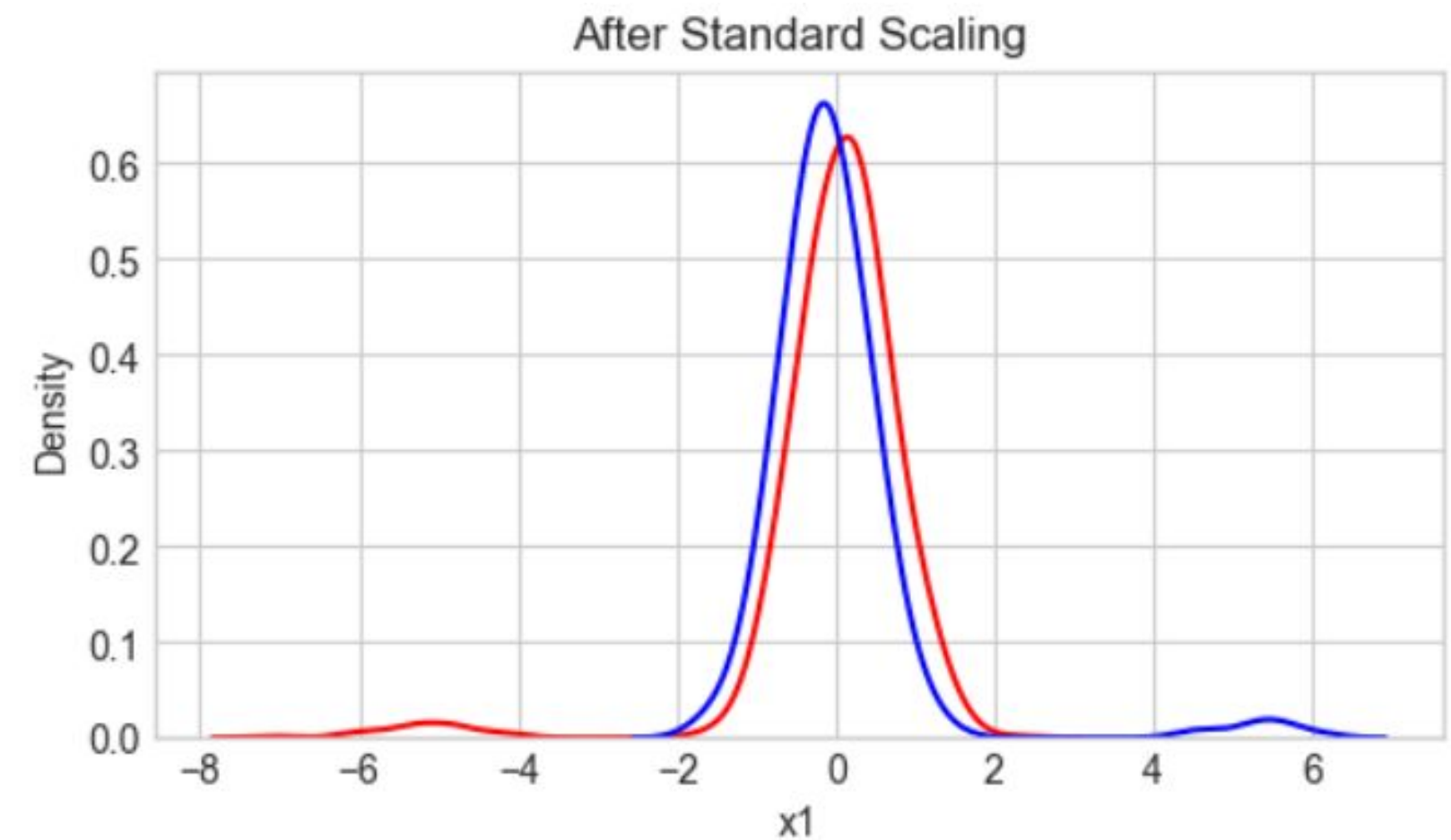
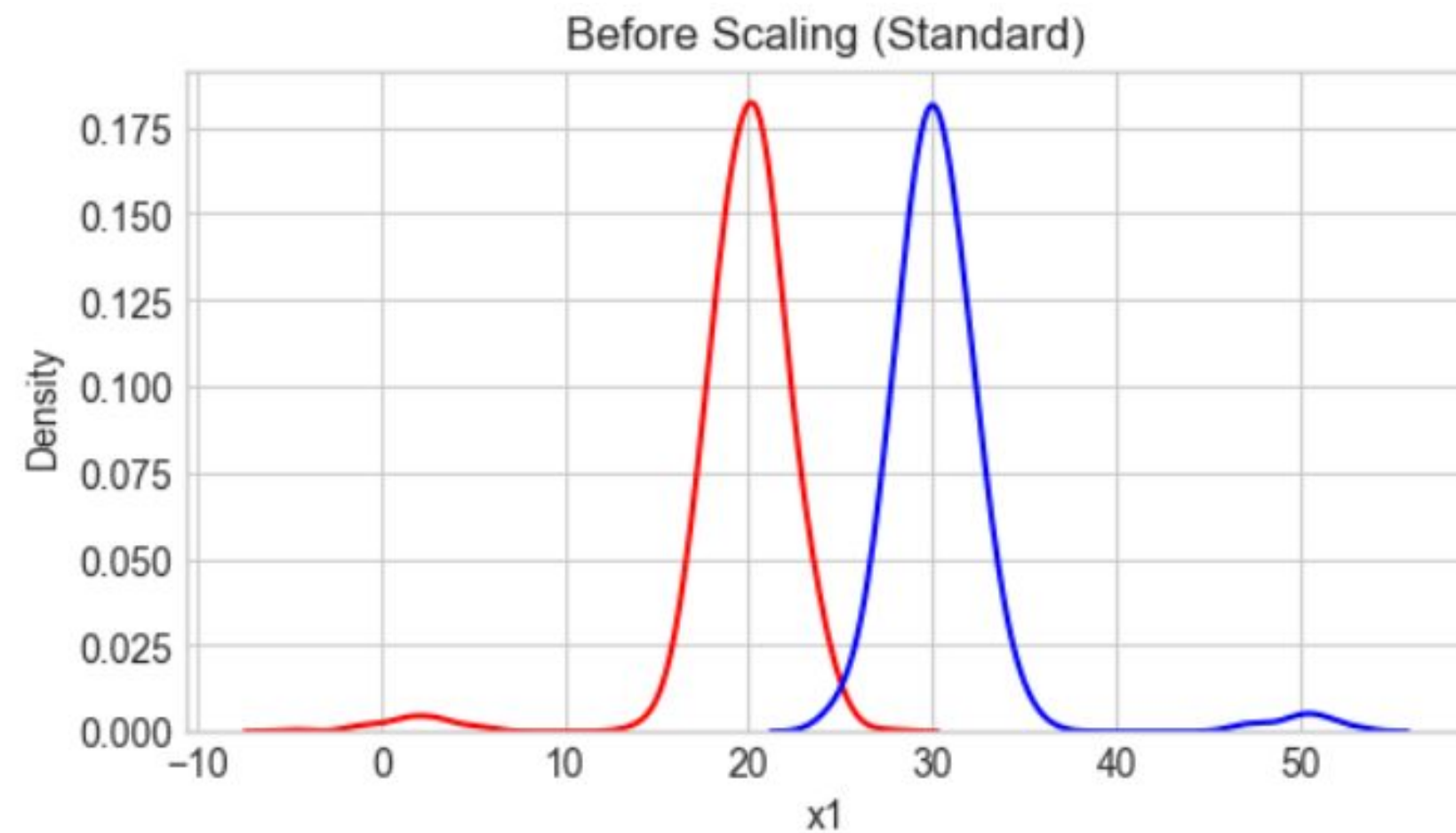
Нормализация данных. MinMaxScaler.



$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Главные шаги препроцессинга

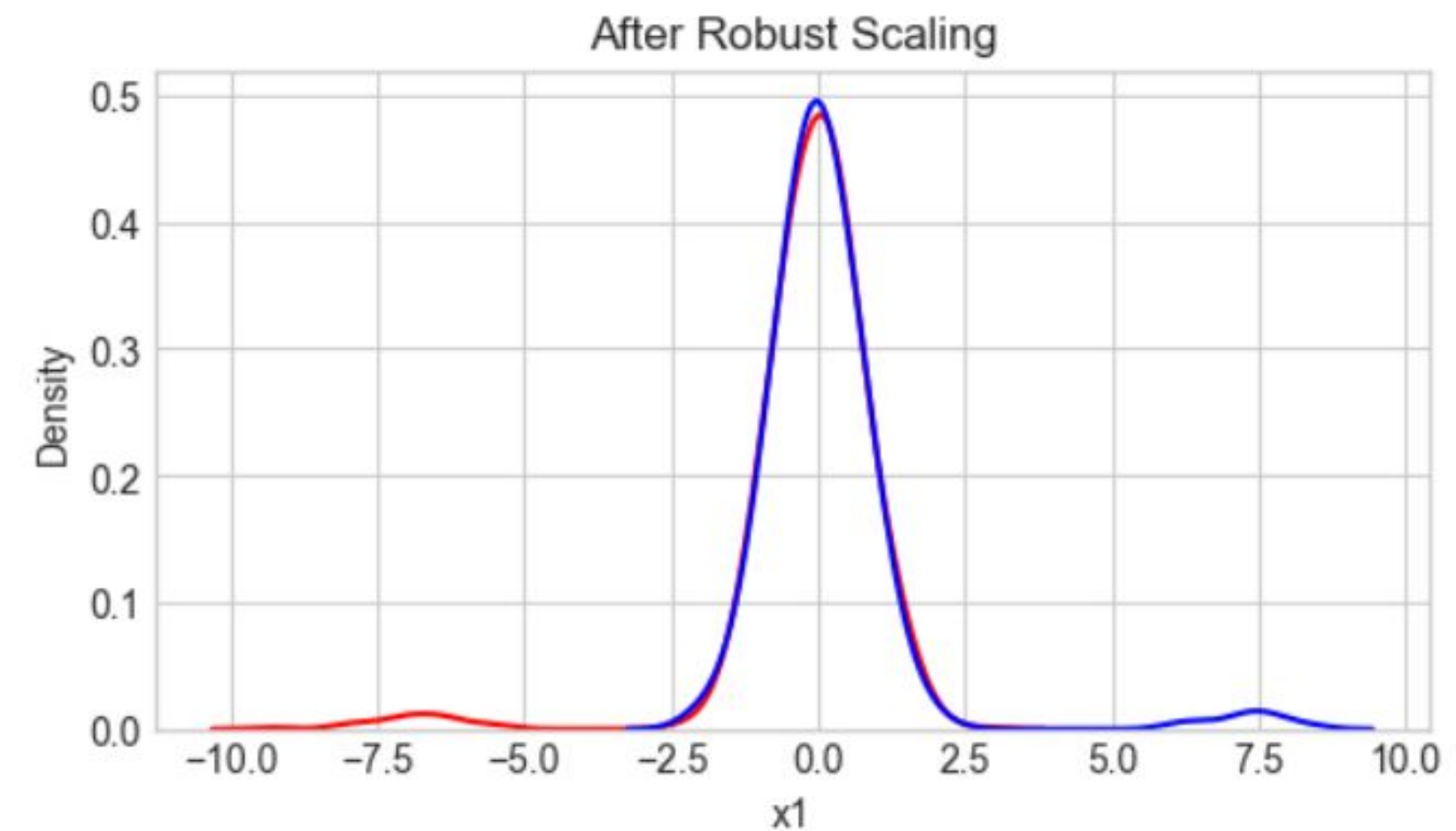
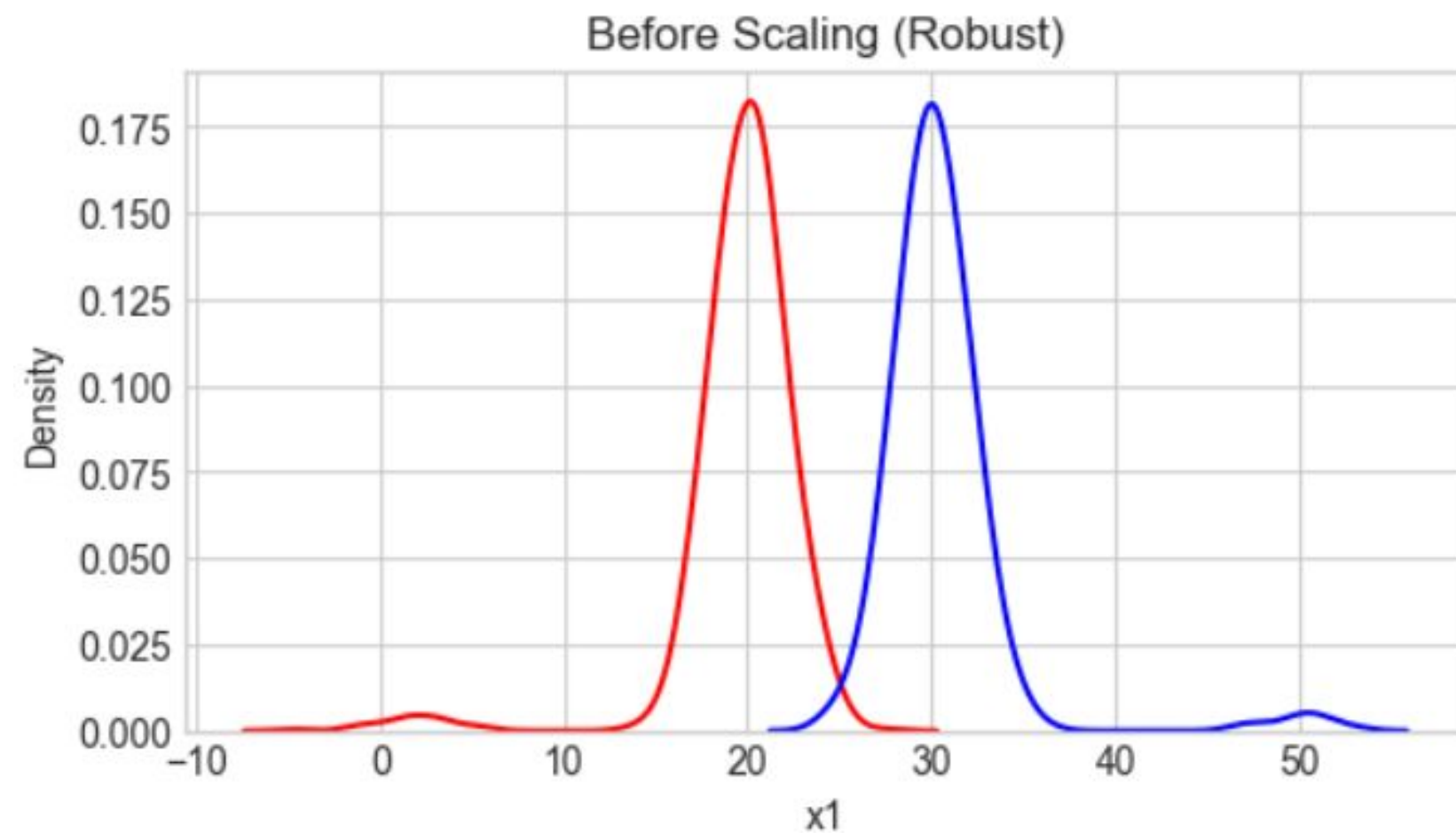
Нормализация данных. StandardScaler.



$$x_{scaled} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

Главные шаги препроцессинга

Нормализация данных. RobustScaler.



$$x_{scaled} = \frac{x - \text{median}(x)}{\text{IQR}(x)}$$

Interquartile Range(IQR) -
разница 25 и 75 квантиля

Главные шаги препроцессинга

Нормализация данных. Выбросы.

Какие из Scaler-ов устойчивы к выбросам?

Главные шаги препроцессинга

Нормализация данных. Выбросы.

Какие из Scaler-ов устойчивы к выбросам?

MinMaxScaler - неустойчив.

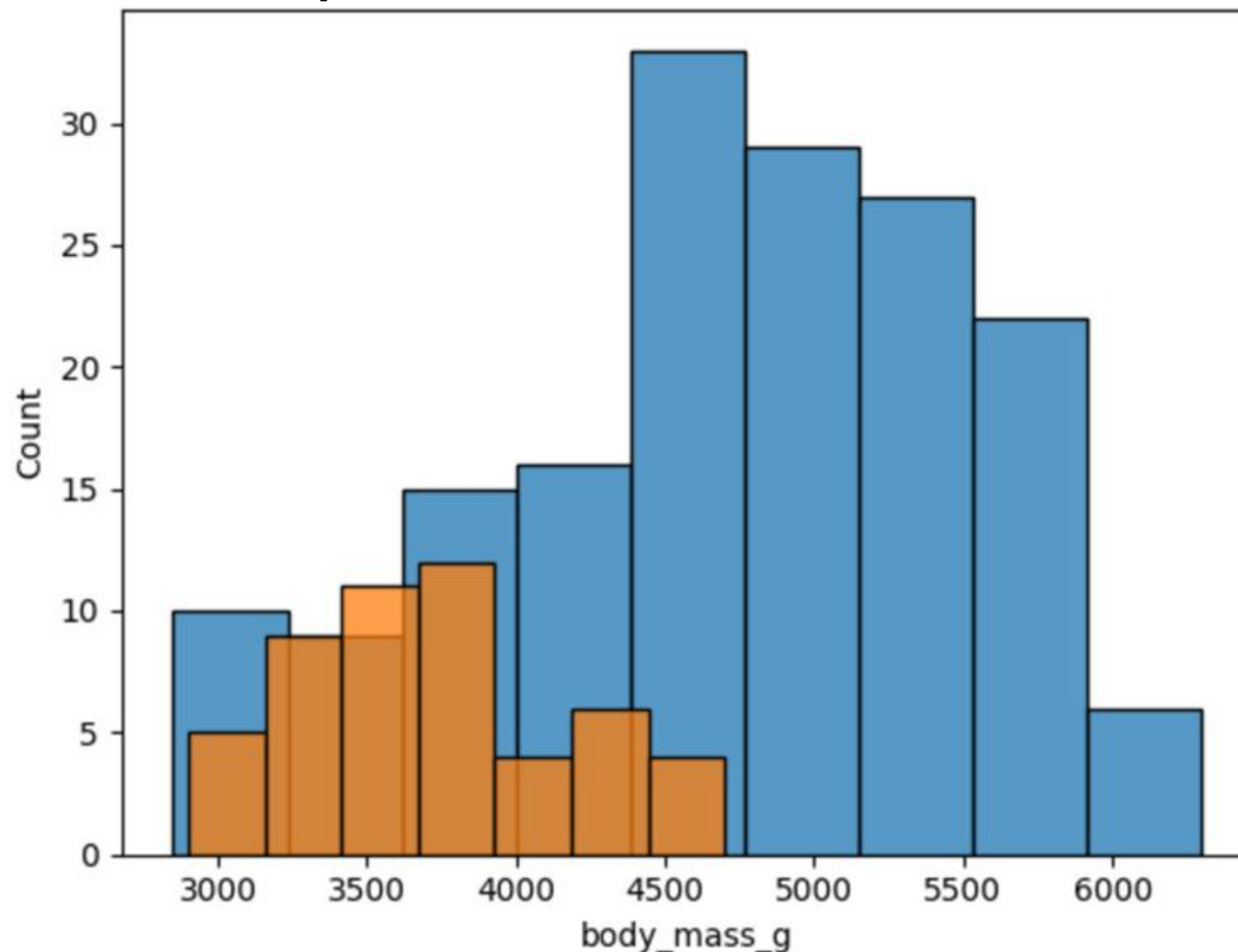
StandardScaler - неустойчив.

RobustScaler - устойчив.

Главные шаги препроцессинга

Визуализация. Гистограммы.

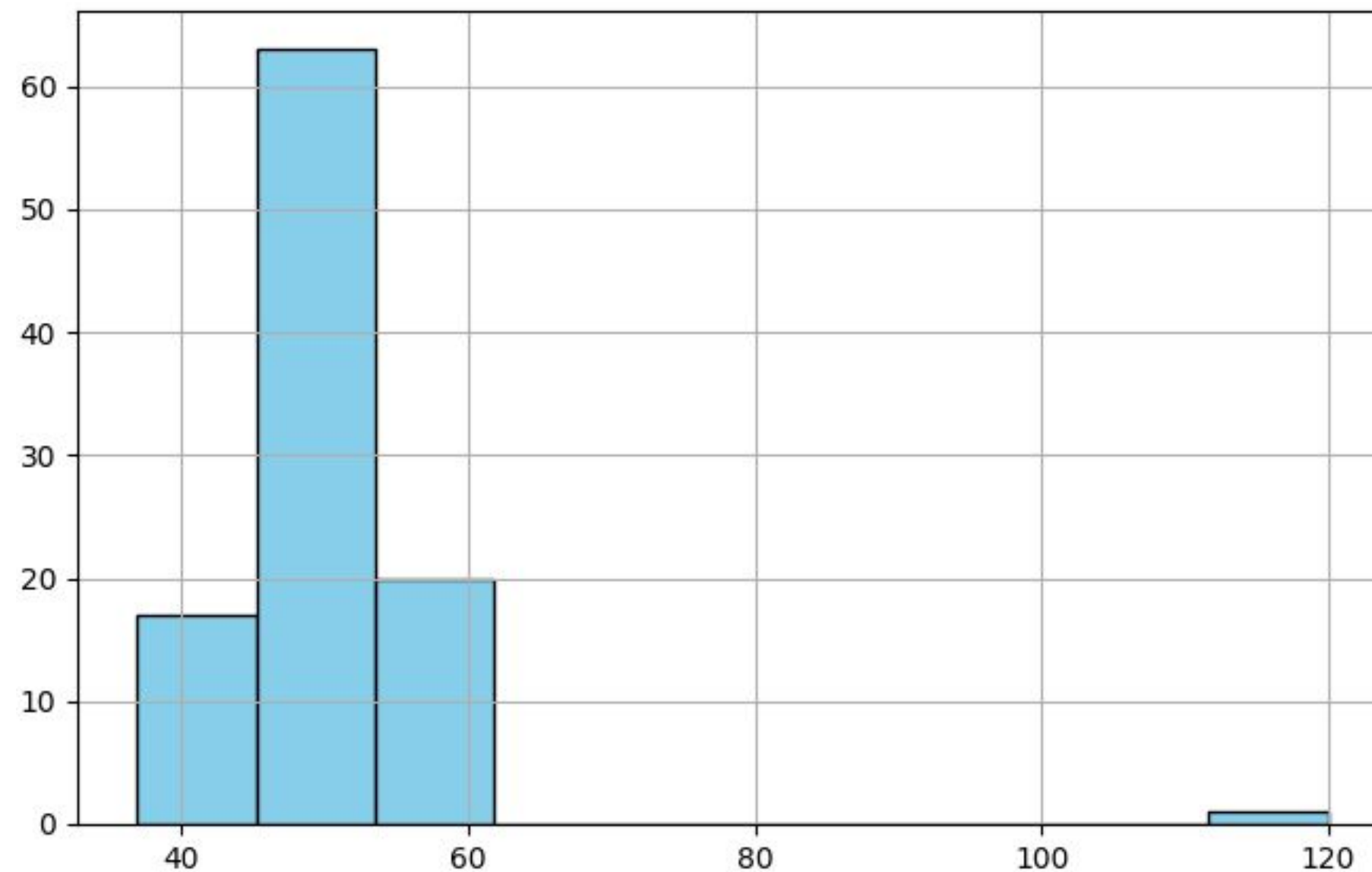
Что не так с гистограммой?



Главные шаги препроцессинга

Визуализация. Гистограммы.

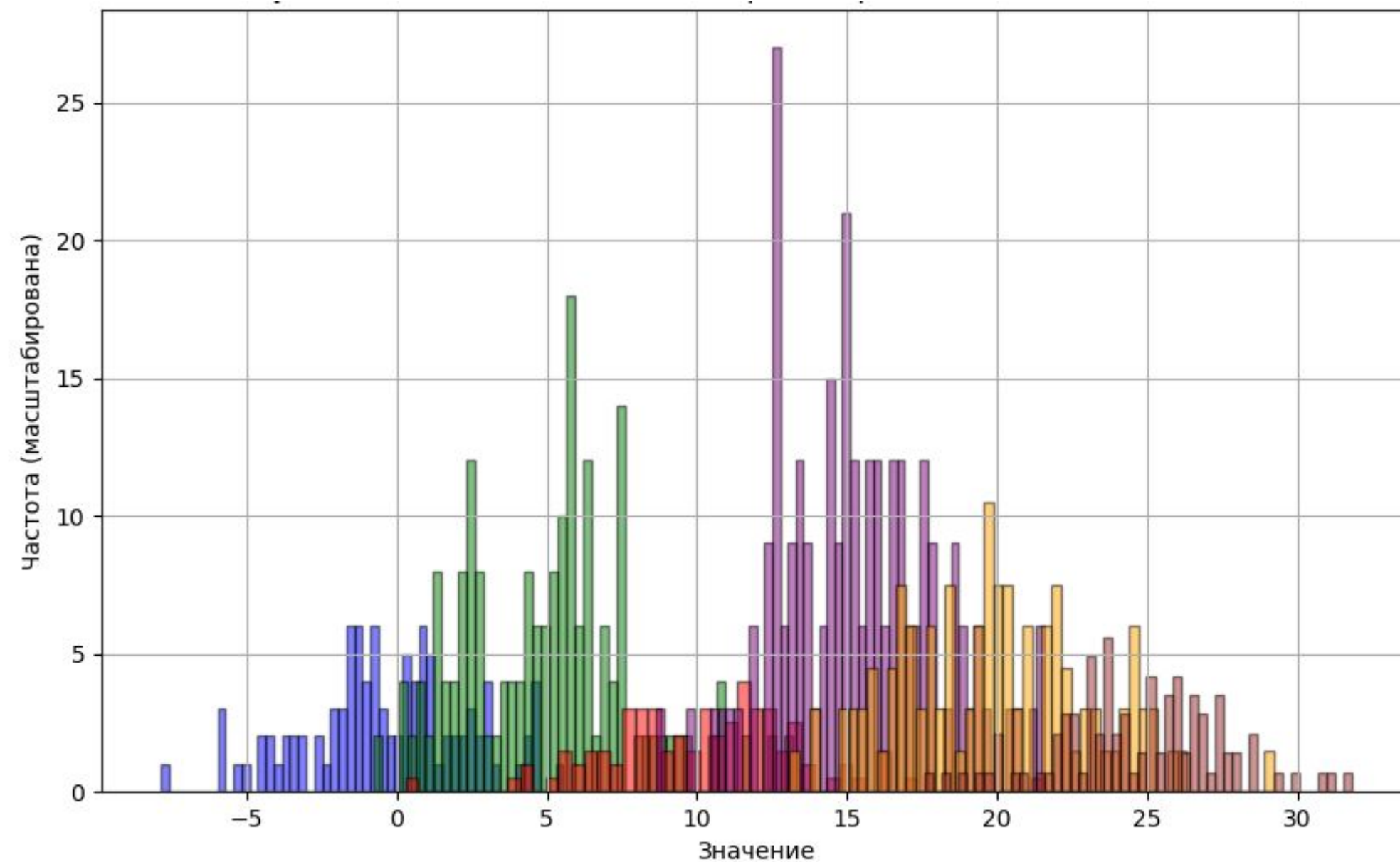
Что не так с гистограммой?



Главные шаги препроцессинга

Визуализация. Гистограммы.

ЧТО НЕ ТАК С ГИСТОГРАММОЙ????????



The histogram displays the frequency of 1000 generated samples. The x-axis represents the value of the samples, ranging from -5 to 30. The y-axis represents the frequency, ranging from 0.0 to 17.5. The distribution is multimodal, with peaks at approximately -2, 0, 2, 5, 8, 12, 15, 20, and 25. The bars are colored in a gradient from blue to red to yellow to orange to brown to pink to purple.

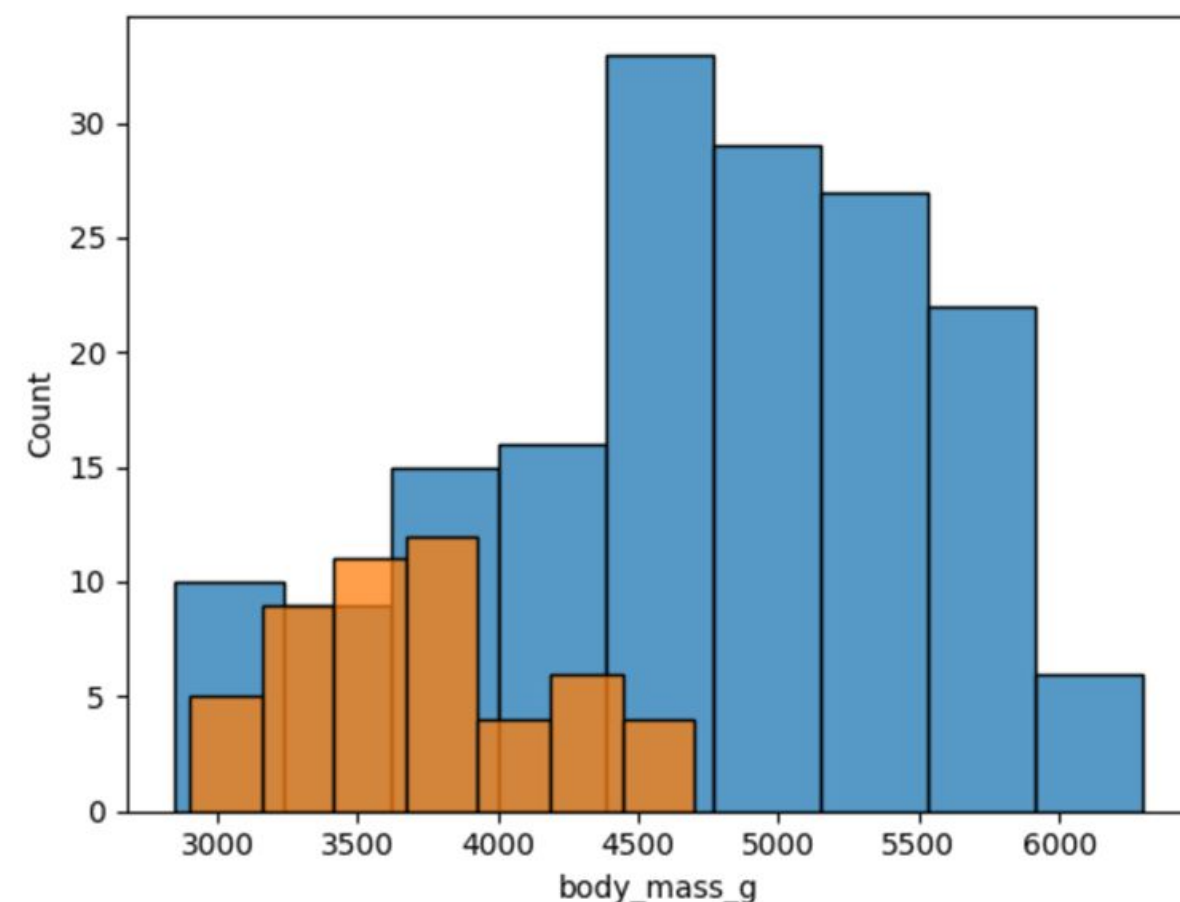
Главные шаги препроцессинга

Визуализация. Гистограммы.

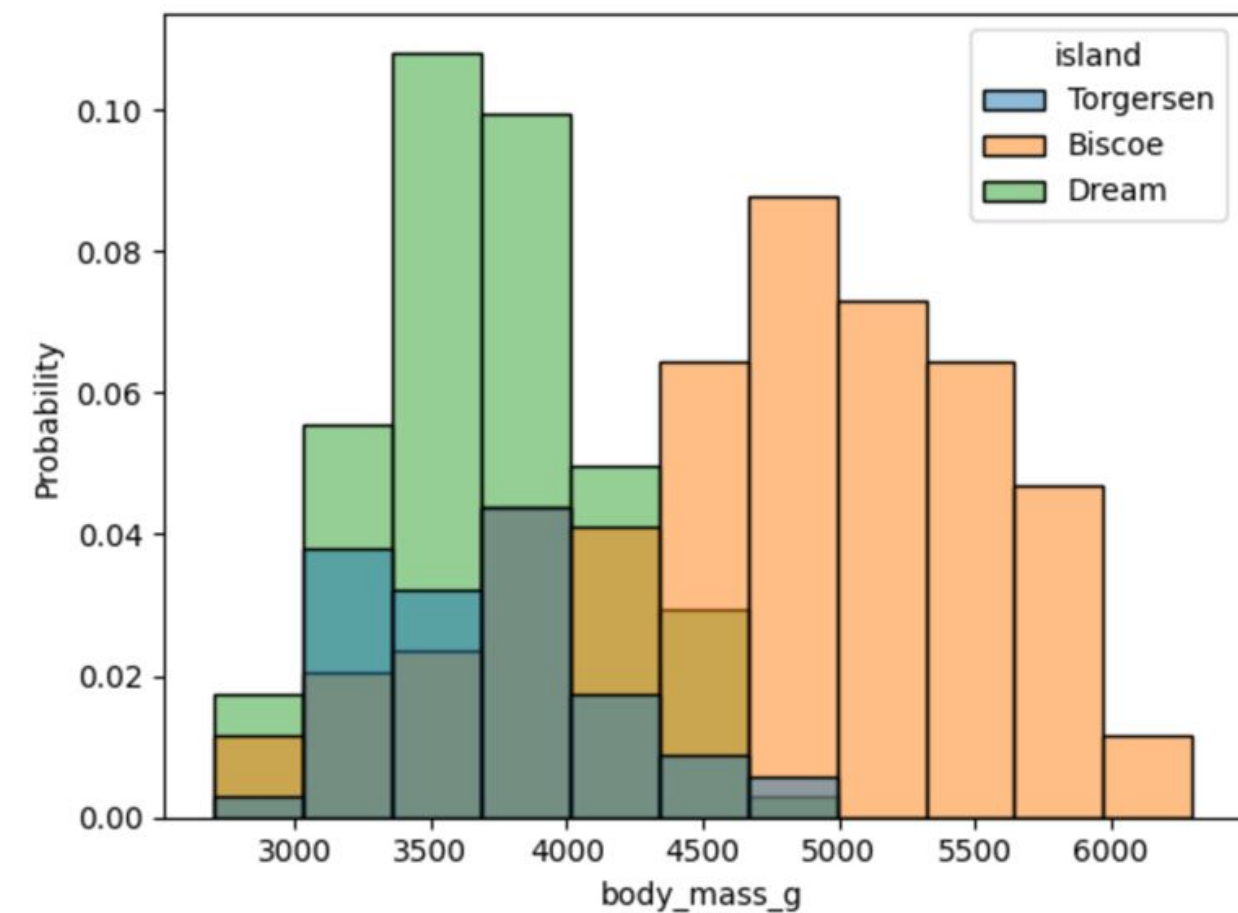
ВАЖНЫЕ правила при изображении гистограмм:

1. Ширина бинов должна быть одинакова.
2. Высота должна быть пропорциональна для каждой части данных.

ПЛОХО

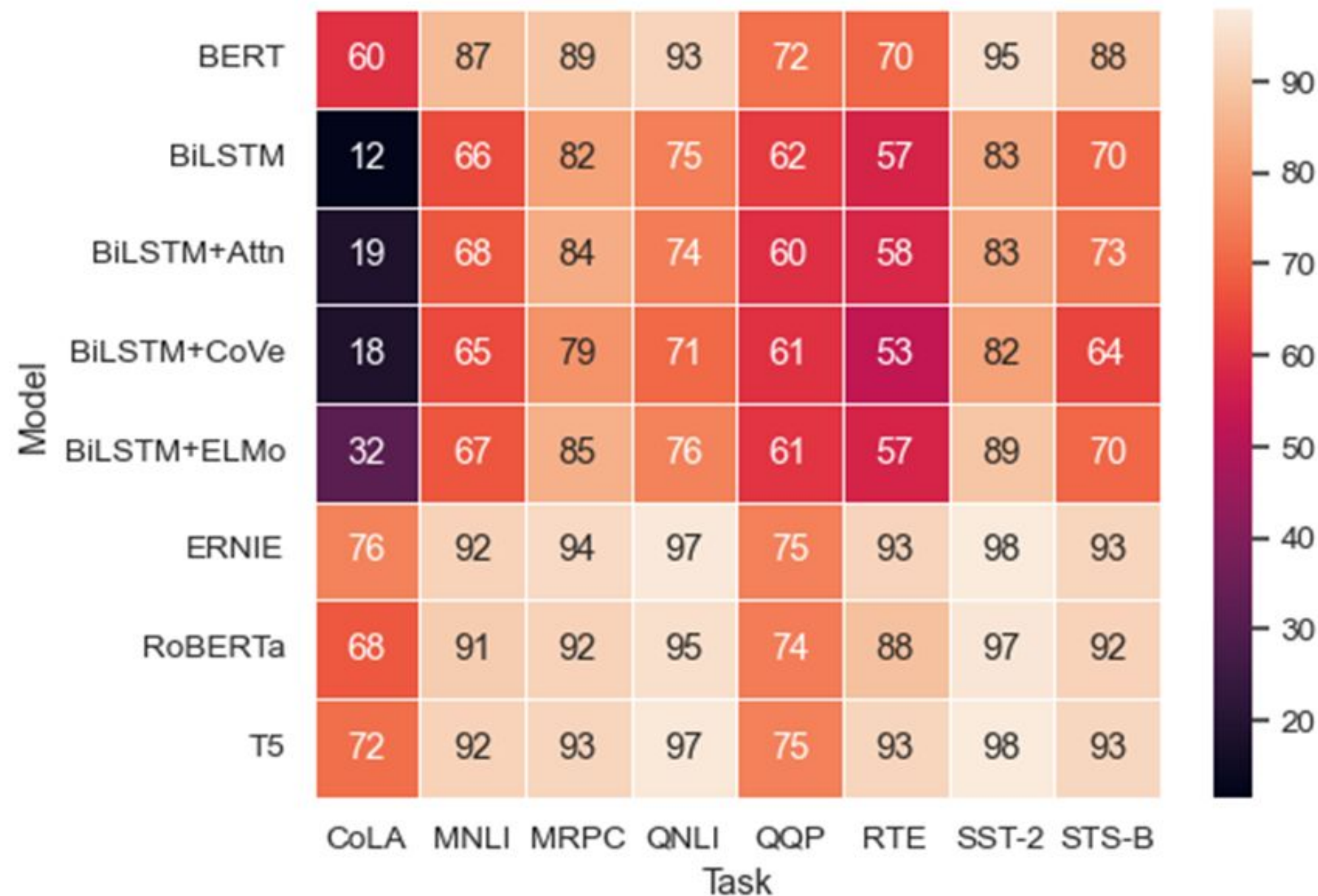


НОРМАЛЬНО



Главные шаги препроцессинга

Визуализация. Heatmap.



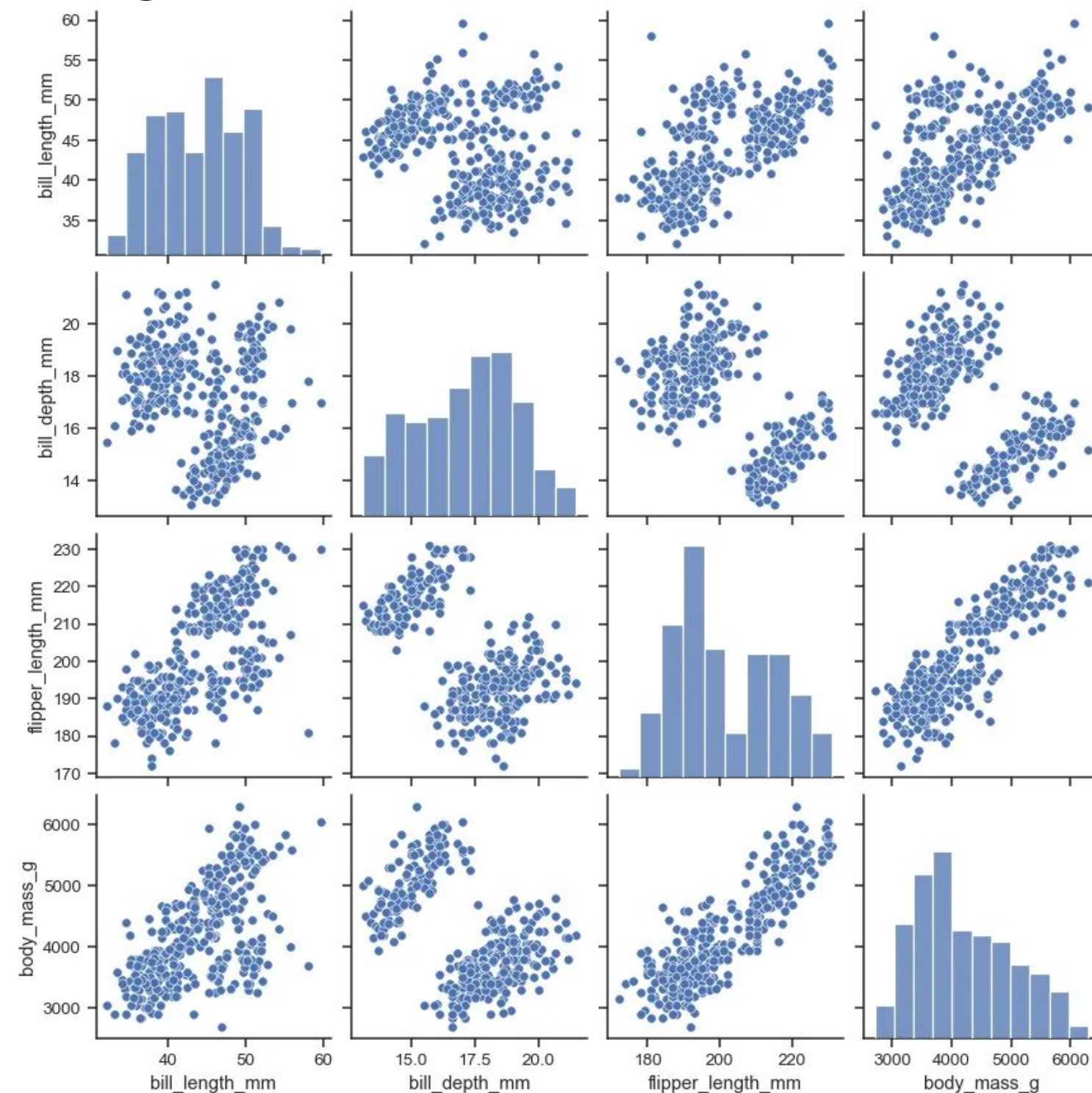
Главные шаги препроцессинга

Визуализация. Heatmap для корреляций.



Главные шаги препроцессинга

Визуализация. Pairplot.



Ко(т)нец презентации



Далее:
Стат. гипотезы

Теоретический минимум

- Функция распределения и плотности
- Ключевые распределения
- Математическое ожидание и дисперсия
- Мода, медиана и размах
- Генеральная совокупность и выборка
- Характеристики выборки
- Работа с пропущенными значениями
- Работа с категориальными переменными
- Нормализация данных
- Визуализации: histplot, scatterplot, boxplot, violinplot, heatmap