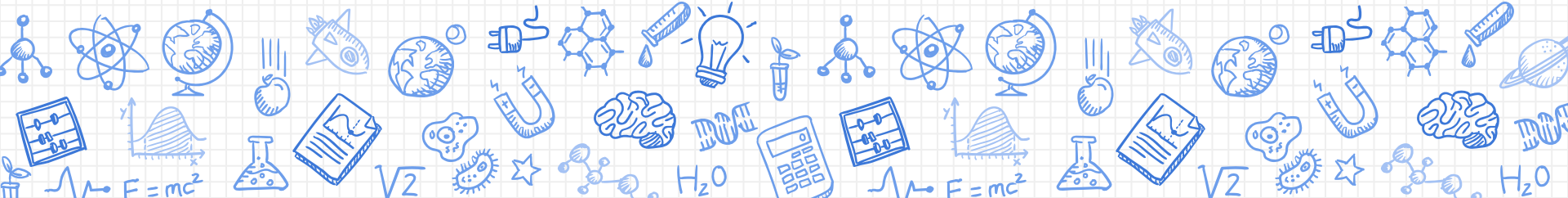
Визуальный анализ данных с Python

Начнем с кусочка статистики



Примеры задач статистики

- ✘ Какова доля брака в продукции завода?
- ✘ Правильно ли настроен станок?
- ✘ Действует ли новое лекарство?
- ✘ Сколько заказать товара на следующий месяц?

Что такое выборка?

Выборка - это часть генеральной совокупности, охватываемая экспериментом. Чаще всего это ограниченный набор чисел или векторов.

Меры среднего

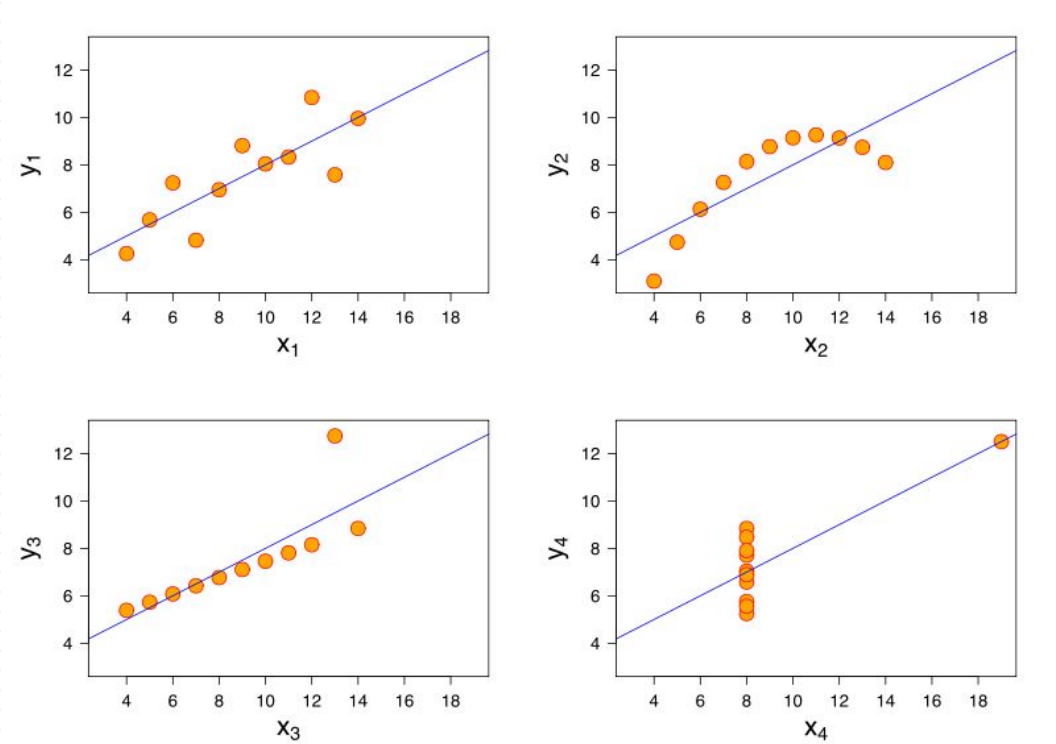
- ✘ Среднее арифметическое
- ✘ Медиана - упорядочиваем числа по возрастанию, берем число посередине ряда.

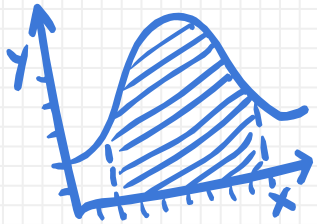
Меры разброса

Дисперсия - чем она больше, тем сильнее разбросаны значения относительно среднего.

$$D = \frac{1}{n}[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2]$$

Квартер Анскомба

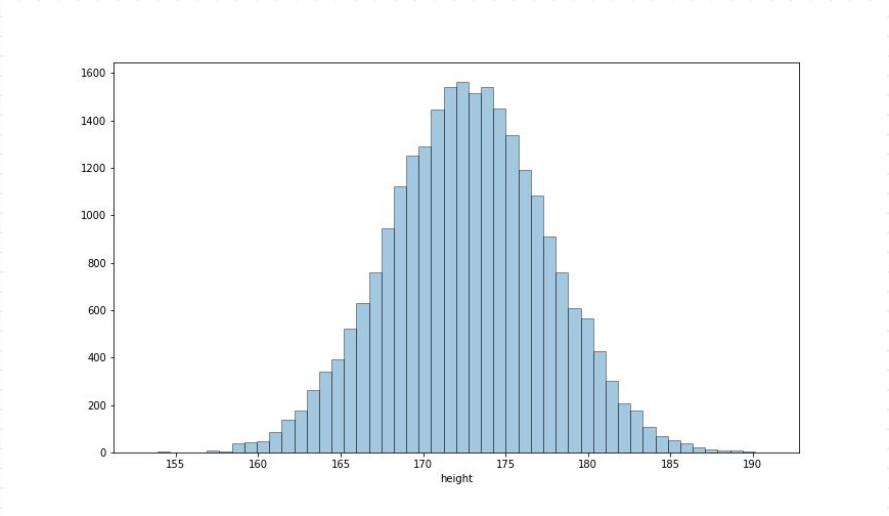




Графики

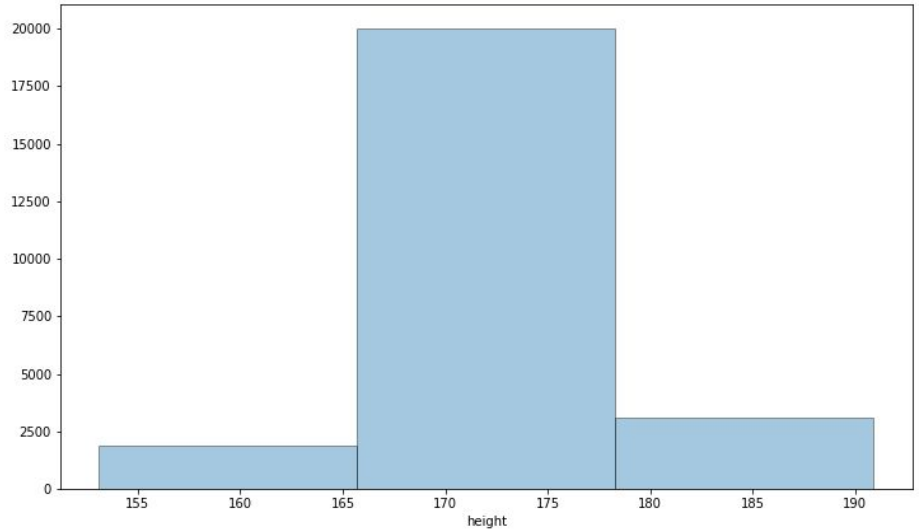
Гистограмма

Разбиваем значения признака на одинаковые промежутки. Для каждого промежутка считаем, сколько точек в него попало.



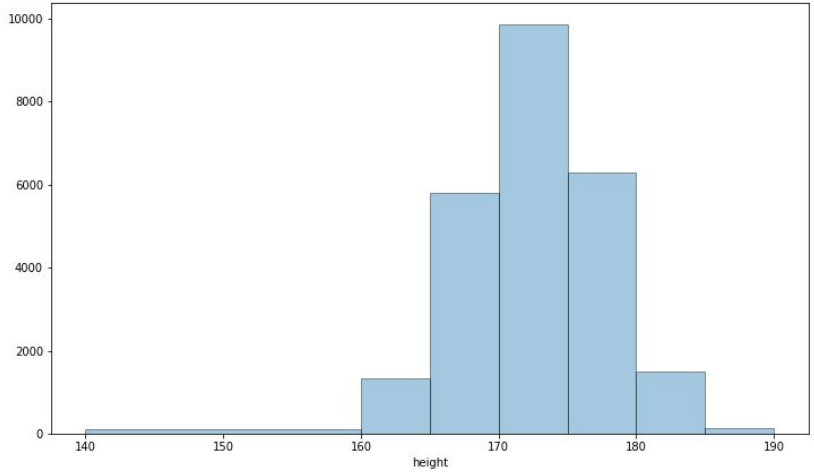
Как выбрать число промежутков?

Плохая гистограмма



Как выбрать число бинов?

Один из вариантов - $\sqrt{N}$



Распределение Гаусса

Параметры:

μ среднее

σ отклонение (корень из дисперсии)