



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

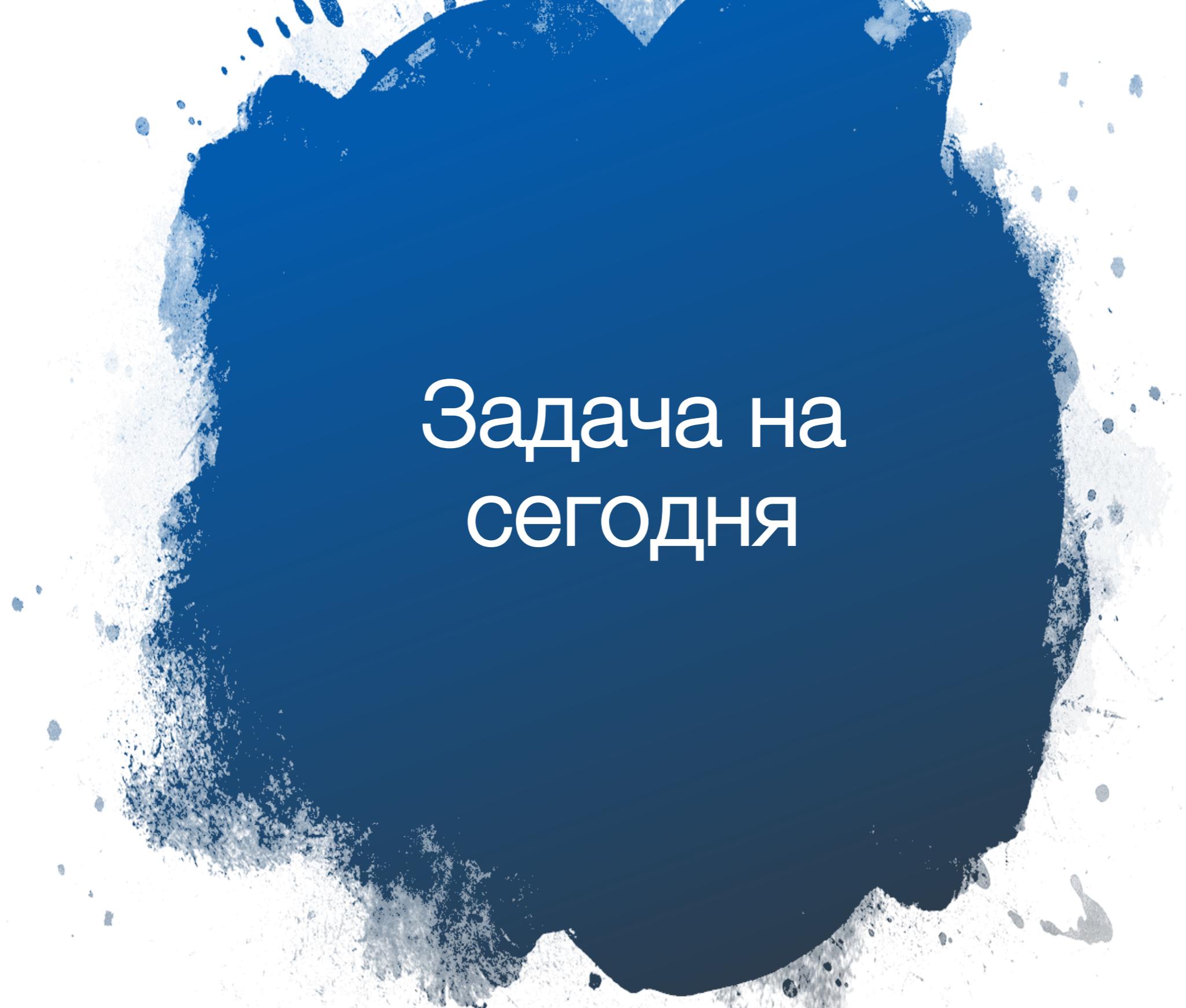
ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ

Все легенды выдуманы, совпадения с реальными фактами
случайны

Теванян Элен

18.04.2019

Москва 2019



Задача на
сегодня

АНАЛИЗИРУЕМ СБОРЫ ФИЛЬМОВ ЗА 2018

- Штатный аналитик КиноПоиска призвал нас в ассистенты и просит дать цифры по сборам фильмов за 2018 год, потому что пиарщики готовят новость для сайта.

КиноПоиск
найди своё кино!

АНАЛИЗИРУЕМ СБОРЫ ФИЛЬМОВ ЗА 2018

- Штатный аналитик КиноПоиска призвал нас в ассистенты и просит дать цифры по сборам фильмов за 2018 год, потому что пиарщики готовят новость для сайта.
- Как будем решать задачу?

КиноПоиск
найди своё кино!

СОБЕРЕМ ДАННЫЕ

КиноПоиск Афиша Фильмы Рейтинги Медиа Онлайн 🔥 Игры

Самые кассовые фильмы

Таблица отражает информацию о самых кассовых фильмах кинопроката за 2018 год.

№	фильм	сборы
1	Мстители: Война бесконечности (2018) Avengers: Infinity War	\$2 048 359 754 9
2	Чёрная Пантера (2018) Black Panther	\$1 346 913 161
3	Мир Юрского периода 2 (2018) Jurassic World: Fallen Kingdom	\$1 309 484 461
4	Суперсемейка 2 (2018) Incredibles 2	\$1 242 805 359
5	Аквамен (2018) Aquaman	\$1 147 661 807
6	Богемская рапсодия (2018) Bohemian Rhapsody	\$901 684 543
7	Веном (2018) Venom	\$855 013 954
8	Миссия невыполнима: Последствия (2018) Mission: Impossible - Fallout	\$791 115 104
9	Дэдпул 2 (2018) Deadpool 2	\$785 046 920
10	Фантастические твари: Преступления Грин-де-Вальда (2018) Fantastic Beasts: The Crimes of Grindelwald	\$653 655 901

НЕМНОГО ОБРАБОТАЕМ

jupyter BoxOffice2018 Last Checkpoint: 30 минут назад (unsaved changes) [Logout](#)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [1]:

```
1 import pandas as pd
2 import numpy as np
```

In [2]:

```
1 data = pd.read_excel('Box_office_2018.xlsx')
2 data.head()
```

Out[2]:

	Название	Сборы
0	Мстители: Война бесконечности (2018)	\$2 048 359 754
1	Чёрная Пантера (2018)	\$1 346 913 161
2	Мир Юрского периода 2 (2018)	\$1 309 484 461
3	Суперсемейка 2 (2018)	\$1 242 805 359
4	Аквамен (2018)	\$1 147 661 807

In [3]:

```
1 data['Сборы'] = data['Сборы'].str.replace('$', '').str.replace('\xa0', '').astype('int')
2 data['Год'] = data['Название'].str[-6:].str.replace('\xa0', '').str.replace(')', '').astype('int')
3 data['Название'] = data['Название'].str[:-8]
4 data = data[data['Год'] == 2018]
```

In [4]:

```
1 data.head()
```

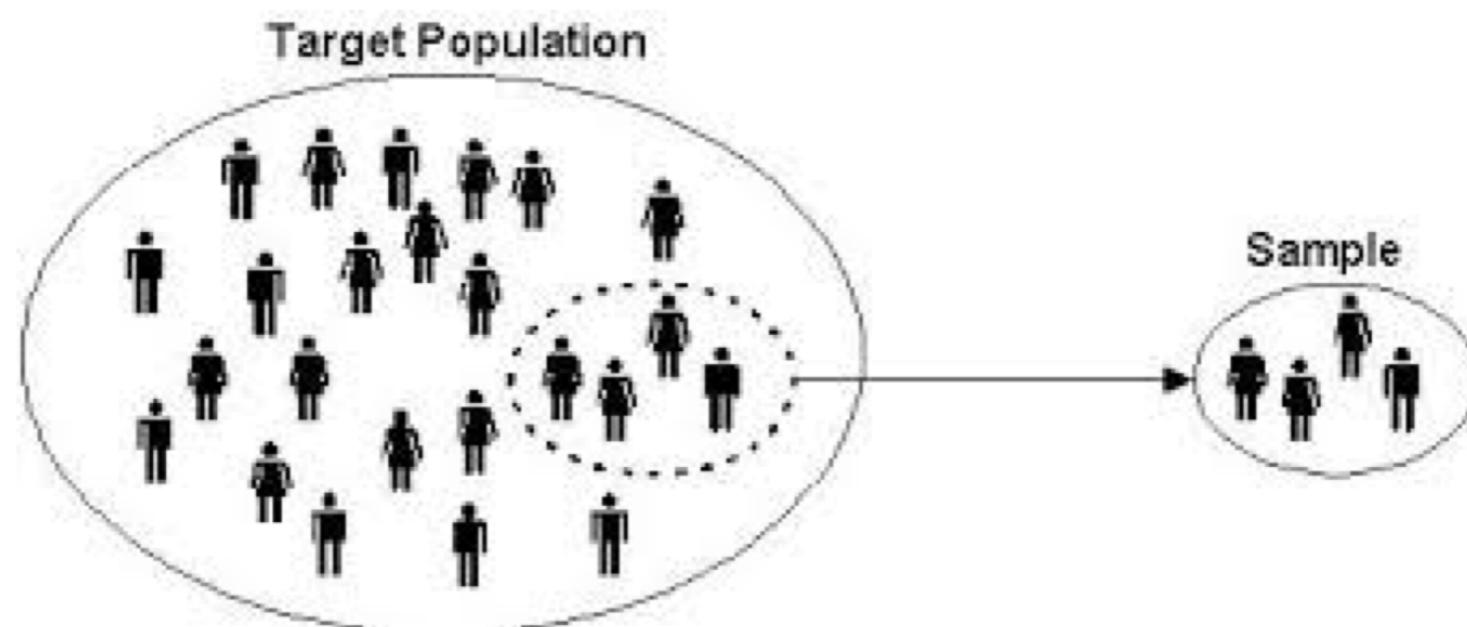
Out[4]:

	Название	Сборы	Год
0	Мстители: Война бесконечности	2048359754	2018
1	Чёрная Пантера	1346913161	2018
2	Мир Юрского периода 2	1309484461	2018
3	Суперсемейка 2	1242805359	2018
4	Аквамен	1147661807	2018

ТО, ЧТО МЫ СОБРАЛИ – ЭТО ВЫБОРКА

Генеральная совокупность – все объекты, про которые будут сделаны выводы, исходя из экспериментов.

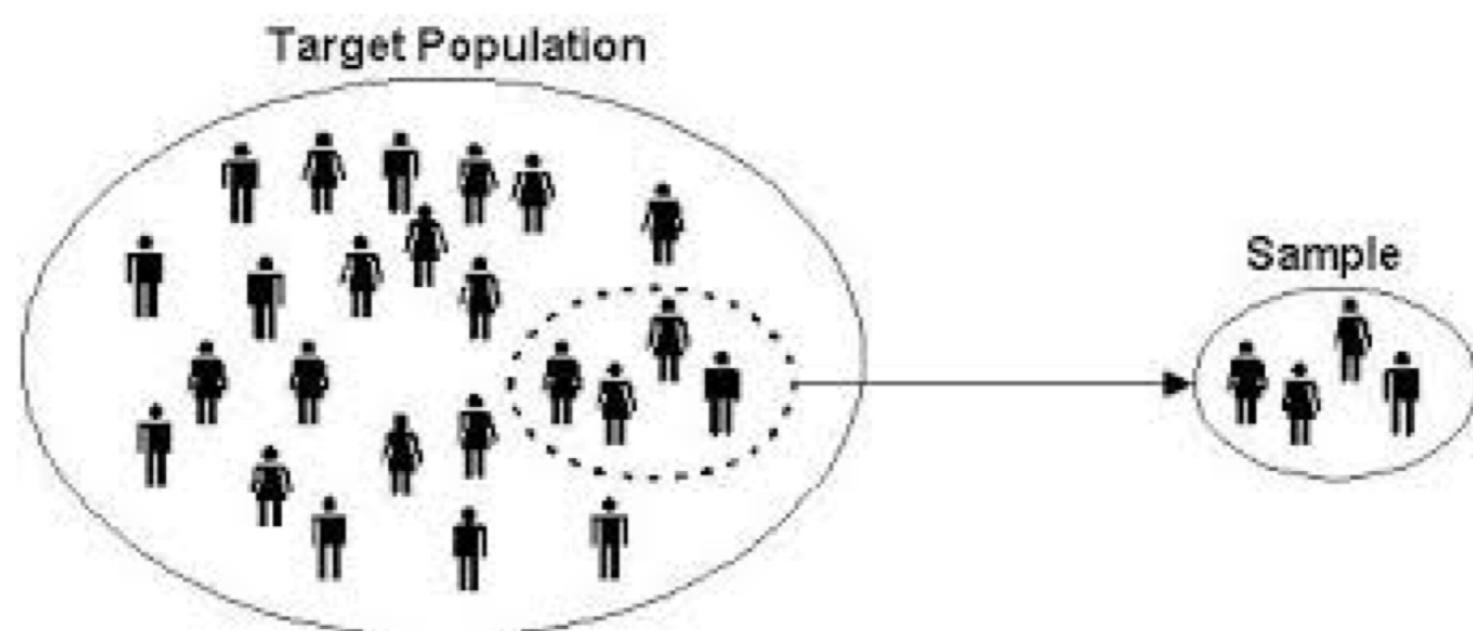
Выборочная совокупность (выборка) – часть генеральной совокупности, которую охватили экспериментом.



ВЫБОРКИ ПОМОГАЮТ ПОЛУЧАТЬ ВЫВОДЫ

- в городе проживает 1 млн человек
- провели соц. опрос об уровне дохода
 - опросили только 2,5 тыс. вместо 1 млн
- опубликовали средний доход по городу

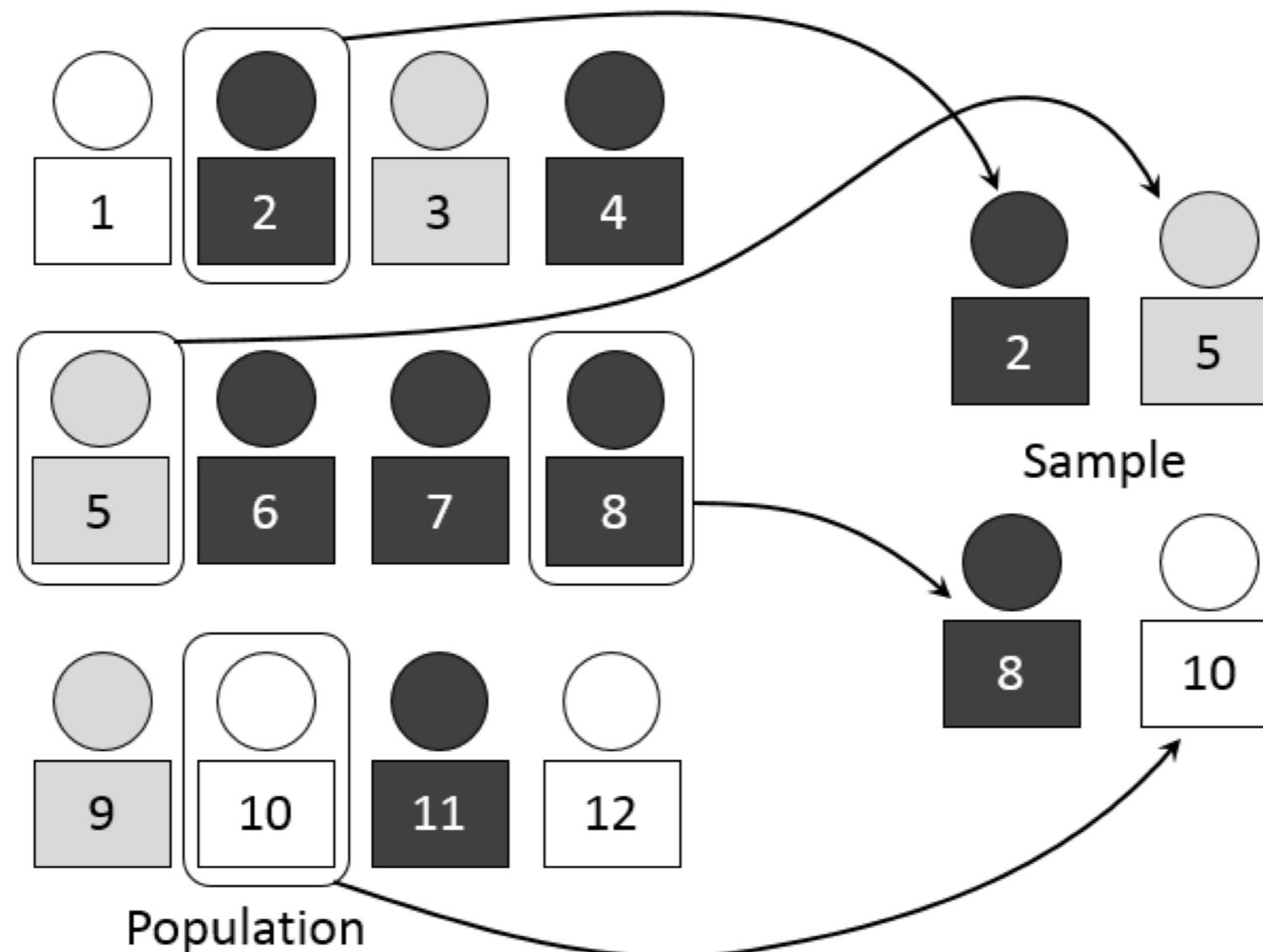
8



ВЫБОРКУ БЫ РЕПРЕЗЕНТАТИВНУЮ

- Репрезентативность выборки определяет, насколько корректно делать выводы обо всей совокупности, опираясь только ⁹ на её подмножество.

КАКАЯ ОНА, РЕПРЕЗЕНТАТИВНАЯ ВЫБОРКА?



А ЗАЧЕМ НАМ ВЫБОРКИ?

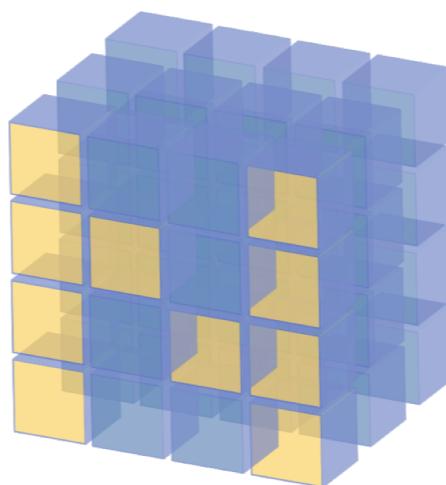
Выборка позволяет сделать выводы о генеральной совокупности:

- средний доход населения
- разброс цен на рынке недвижимости
- процент людей, получивших высшее образование
- средний возраст ¹¹ подписчиков Netflix
- порог зачисления на майнер

СЕГОДНЯ НАМ ПОМОГАЮТ



Pandas



NumPy

Описательные статистики

СТАТИСТИКА – ЭТО

Статистика – это функция от выборки.

КОЛИЧЕСТВО ЭЛЕМЕНТОВ

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Сколько в ней элементов?

КОЛИЧЕСТВО ЭЛЕМЕНТОВ

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Сколько в ней элементов?

10

КОЛИЧЕСТВО ЭЛЕМЕНТОВ

Количество элементов

```
In [6]: 1 data[['Сборы']].count()
```

```
Out[6]: Сборы    278  
         dtype: int64
```

```
In [7]: 1 np.size(data['Сборы'])
```

```
Out[7]: 278
```

МИНИМУМ И МАКСИМУМ

$$f(X) = \min(X) = \min\{x_1, \dots, x_n\}$$

$$f(X) = \max(X) = \max\{x_1, \dots, x_n\}$$

СЧИТАЕМ МИНИМУМ И МАКСИМУМ

$$f(X) = \min(X) = \min\{x_1, \dots, x_n\}$$

$$f(X) = \max(X) = \max\{x_1, \dots, x_n\}$$

19

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каковы минимум и максимум?

СЧИТАЕМ МИНИМУМ И МАКСИМУМ

$$f(X) = \min(X) = \min\{x_1, \dots, x_n\}$$

$$f(X) = \max(X) = \max\{x_1, \dots, x_n\}$$

20

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

$$\min(X) = -10 \quad \max(X) = 5$$

СЧИТАЕМ МИНИМУМ

Минимум

```
In [8]: 1 data[ [ 'Сборы' ] ].min()
```

```
Out[8]: Сборы      528  
         dtype: int64
```

```
In [9]: 1 np.min(data[ 'Сборы' ])
```

```
Out[9]: 528
```

СЧИТАЕМ МИНИМУМ

```
In [28]: 1 data[data['Сборы'] == data['Сборы'].min()]
```

Out[28]:

	Название	Сборы	Год
484	Высшая сила	528	2018

СЧИТАЕМ МАКСИМУМ

Максимум

```
In [10]: 1 data[ [ 'Сборы' ] ].max( )
```

```
Out[10]: Сборы      2048359754
          dtype: int64
```

```
In [11]: 1 np.max(data[ 'Сборы' ])
```

```
Out[11]: 2048359754
```

СЧИТАЕМ МАКСИМУМ

```
In [29]: 1 data[data[ 'Сборы' ] == data[ 'Сборы' ].max( )]
```

Out[29]:

	Название	Сборы	Год
0	Мстители: Война бесконечности	2048359754	2018

ВЫБОРОЧНОЕ СРЕДНЕЕ

$$f(X) = \bar{X} = \frac{x_1 + \dots + x_n}{n}$$

СЧИТАЕМ ВЫБОРОЧНОЕ СРЕДНЕЕ

$$f(X) = \bar{X} = \frac{x_1 + \dots + x_n}{n}$$

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каково выборочное среднее?

СЧИТАЕМ ВЫБОРОЧНОЕ СРЕДНЕЕ

$$f(X) = \bar{X} = \frac{x_1 + \dots + x_n}{n}$$

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каково выборочное среднее?

$$\bar{X} = \frac{1 + 5 + 3 + (-1) + (-4) + 3 + 3 + (-10) + 2 + (-1)}{10} = 0.1$$

СЧИТАЕМ ВЫБОРОЧНОЕ СРЕДНЕЕ

Среднее

```
In [12]: 1 data[['Сборы']].mean()
```

```
Out[12]: Сборы      1.101081e+08
          dtype: float64
```

```
In [13]: 1 np.mean(data['Сборы'])
```

```
Out[13]: 110108083.57913668
```

ВЫБОРОЧНАЯ ДИСПЕРСИЯ

$$f(X) = s^2(X) = \frac{(x_1 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n}$$

СЧИТАЕМ ВЫБОРОЧНУЮ ДИСПЕРСИЮ

$$f(X) = s^2(X) = \frac{(x_1 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n}$$

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Какова выборочная дисперсия?

СЧИТАЕМ ВЫБОРОЧНУЮ ДИСПЕРСИЮ

$$f(X) = s^2(X) = \frac{(x_1 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n}$$

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Какова выборочная дисперсия?

$$s^2(X) = \frac{(1 - 0.1)^2 + (5 - 0.1)^2 + \dots + (-1 - 0.1)^2}{10} = 17.49$$

СЧИТАЕМ ВЫБОРОЧНУЮ ДИСПЕРСИЮ

Дисперсия

```
In [14]: 1 data[ [ 'Сборы' ] ].var()
```

```
Out[14]: Сборы      5.836131e+16
          dtype: float64
```

```
In [15]: 1 np.var(data[ 'Сборы' ])
```

```
Out[15]: 5.815137600079342e+16
```

СТАНДАРТНОЕ ОТКЛОНЕНИЕ

$$s(X) = \sqrt{s^2(X)}$$

СЧИТАЕМ СТАНДАРТНОЕ ОТКЛОНЕНИЕ

$$s(X) = \sqrt{s^2(X)}$$

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каково стандартное отклонение?

$$s^2(X) = 17.49$$

СЧИТАЕМ СТАНДАРТНОЕ ОТКЛОНЕНИЕ

$$s(X) = \sqrt{s^2(X)}$$

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каково стандартное отклонение?

$$s^2(X) = 17.49$$

$$s(X) = \sqrt{s^2(X)} \approx 4.182$$

СЧИТАЕМ СТАНДАРТНОЕ ОТКЛОНЕНИЕ

Стандартное отклонение

```
In [16]: 1 data[ [ 'Сборы' ] ].std()
```

```
Out[16]: Сборы      2.415809e+08  
          dtype: float64
```

```
In [17]: 1 np.std(data[ 'Сборы' ])
```

```
Out[17]: 241145964.09808195
```

ПЕРЦЕНТИЛИ ПОРЯДКА К

Перцентиль порядка К – такое число, что К значений выборки меньше этого числа.

ПЕРЦЕНТИЛИ ПОРЯДКА К

- Проще всего вычислять по упорядоченному (отсортированному) набору чисел.
- Далее предполагаем, что выборка упорядочена:

$$x_1 < \dots < x_n$$

ПЕРЦЕНТИЛИ ПОРЯДКА K: ИЗВЕСТНЫЕ

- Нижняя квартиль: $K = 25$

$$LQ(X) = x[0.25 \cdot (n+1)]$$

- медиана (вторая квартиль): $K = 50$

$$M(X) = x[0.5 \cdot (n+1)]$$

- верхняя квартиль: $K = 75$

$$UQ(X) = x[0.75 \cdot (n+1)]$$

СЧИТАЕМ КВАРТИЛИ

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каковы значения LQ, M и UQ?

СЧИТАЕМ КВАРТИЛИ

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каковы значения LQ, M и UQ?

Сначала упорядочим её: -10, -4, -1, -1, 1, 2, 3, 3, 3, 5

СЧИТАЕМ КВАРТИЛИ

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каковы значения LQ, M и UQ?

Сначала упорядочим её⁴²: -10, -4, -1, -1, 1, 2, 3, 3, 3, 5

- Нижняя квартиль: расположена между 2 и 3 наблюдениями

$$0.25 \cdot (10 + 1) = 2.75$$

$$LQ = \frac{-4 + (-1)}{2} = -2.5$$

СЧИТАЕМ КВАРТИЛИ

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каковы значения LQ, M и UQ?

Сначала упорядочим её⁴³: -10, -4, -1, -1, 1, 2, 3, 3, 3, 5

- верхняя квартиль: расположена между 8 и 9 наблюдениями

$$0.75 \cdot (10 + 1) = 8.25$$

$$UQ = \frac{3 + 3}{2} = 3$$

СЧИТАЕМ КВАРТИЛИ

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Каковы значения LQ, M и UQ?

Сначала упорядочим её⁴⁴: -10, -4, -1, -1, 1, 2, 3, 3, 3, 5

- медиана: расположена между 5 и 6 наблюдениями

$$0.5 \cdot (10 + 1) = 5.5$$

$$M = \frac{1 + 2}{2} = 1.5$$

СЧИТАЕМ КВАРТИЛИ: НИЖНЯЯ

Перцентили

```
In [18]: 1 data[ [ 'Сборы' ] ].quantile(0.25)
```

```
Out[18]: Сборы      1644194.5  
          Name: 0.25, dtype: float64
```

```
In [19]: 1 np.quantile(data[ 'Сборы' ], q=0.25)
```

```
Out[19]: 1644194.5
```

СЧИТАЕМ КВАРТИЛИ: ВЕРХНЯЯ

```
In [20]: 1 data[ [ 'Сборы' ] ].quantile(0.75)
```

```
Out[20]: Сборы    93139428.25  
          Name: 0.75, dtype: float64
```

```
In [21]: 1 np.quantile(data[ 'Сборы' ], q=0.75)
```

```
Out[21]: 93139428.25
```

МЕДИАНА

Медиана

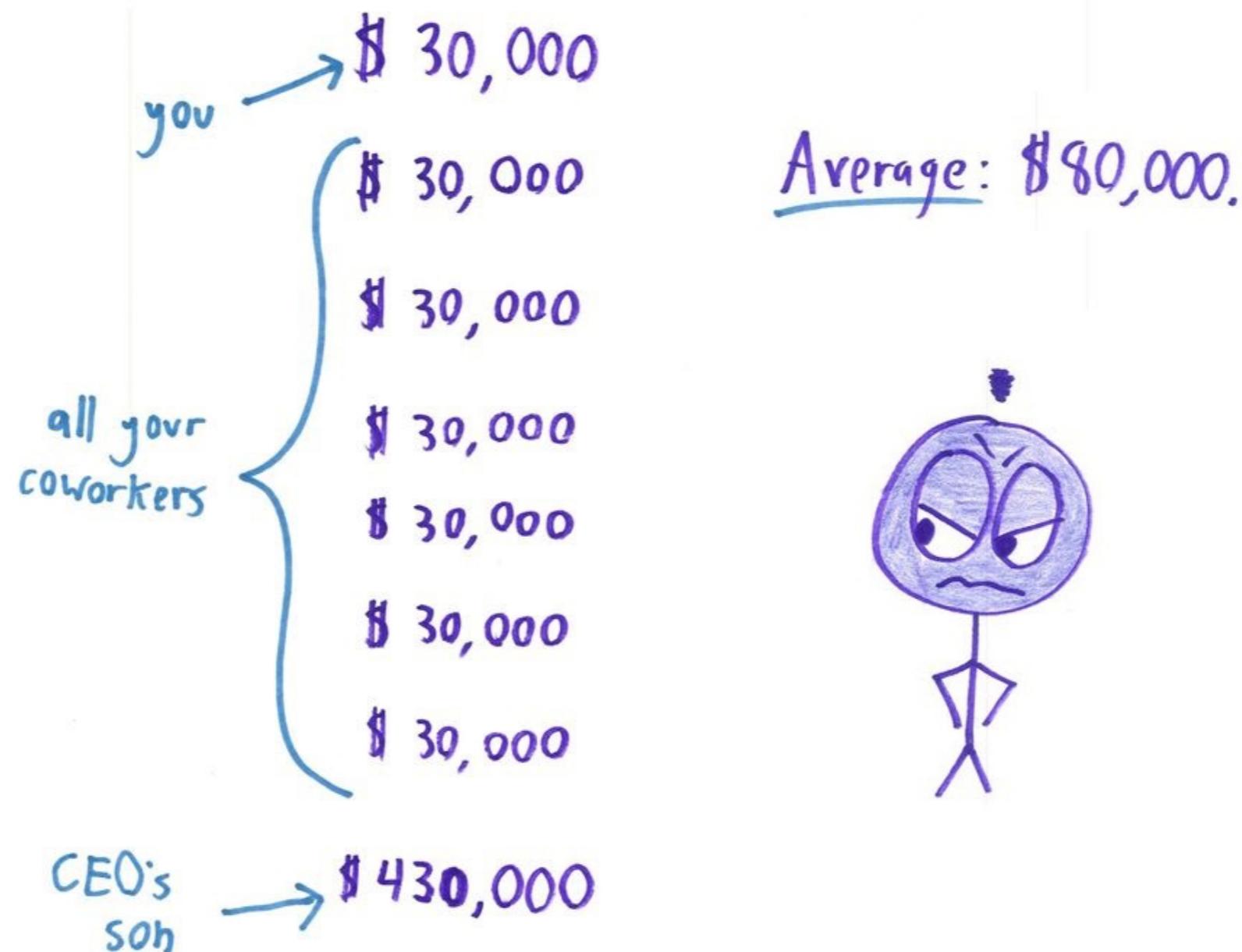
```
In [22]: 1 data[['Сборы']].median()
```

```
Out[22]: Сборы    17552597.5
          dtype: float64
```

```
In [23]: 1 np.median(data['Сборы'])
```

```
Out[23]: 17552597.5
```

СРЕДНЕЕ И МЕДИАНА



ДИСПЕРСИЯ VS КВАРТИЛИ

- Опросили 100 человек
- 99 имеют доход 10.000 рублей
- 1 имеет доход 1.000.000 рублей

ДИСПЕРСИЯ VS КВАРТИЛИ

- Опросили 100 человек
- 99 имеют доход 10.000 рублей
- 1 имеет доход 1.000.000 рублей
- Стандартное отклонение: ~98503

МОДА

- Самое частое значение в выборке.

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Какова мода?

СЧИТАЕМ МОДУ

Дана числовая выборка: 1, 5, 3, -1, -4, 3, 3, -10, 2, -1

Наблюдение	Частота
-10	1
-4	1
-1	2
1	1
2	1
3	3
5	1

СЧИТАЕМ МОДУ

Мода

```
In [24]: 1 data[['Сборы']].mode(axis=1).iloc[0,:]
```

```
Out[24]: 0    2048359754  
Name: 0, dtype: int64
```

```
In [25]: 1 stats.mode(data['Сборы'])
```

```
Out[25]: ModeResult(mode=array([528]), count=array([1]))
```

ВСЕ СТАТИСТИКИ

Все статистики

```
In [26]: 1 data[ [ 'Сборы' ] ].describe()
```

Out[26]:

Сборы	
count	2.780000e+02
mean	1.101081e+08
std	2.415809e+08
min	5.280000e+02
25%	1.644194e+06
50%	1.755260e+07
75%	9.313943e+07
max	2.048360e+09

```
In [27]: 1 stats.describe(data[ 'Сборы' ])
```

Out[27]: DescribeResult(nobs=278, minmax=(528, 2048359754), mean=110108083.579136
=4.103816836615013, kurtosis=21.4290706719513)

ТЕПЕРЬ МЫ УМЕЕМ ОПИСЫВАТЬ ДАННЫЕ

Научились считать следующие статистики:

- минимум / максимум
- среднее
- дисперсия
- стандартное отклонение
- мода
- перцентили различных порядков
(популярны 25%, 50% и 75%)



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ