

ЧТО "НЕ ТАК"?



ML Basic

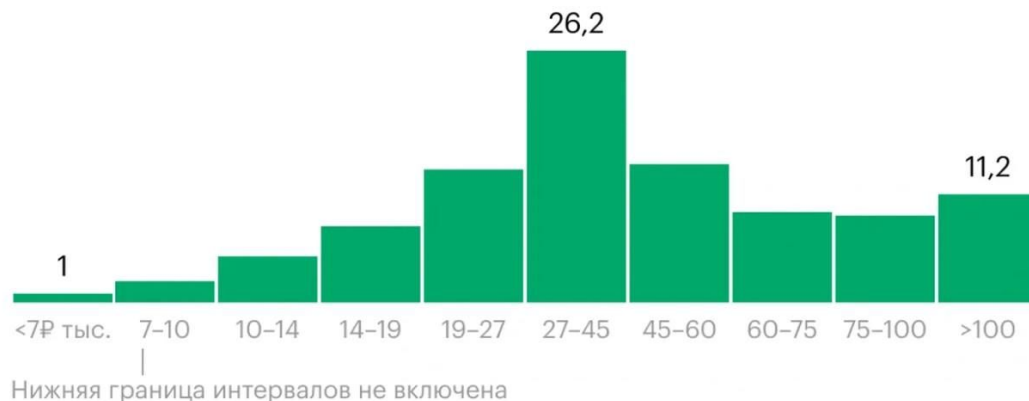
Визуализация

Начнем в 20:01 мск

otus.ru

Распределение населения России по величине среднедушевых доходов

Данные за 2023 год, %

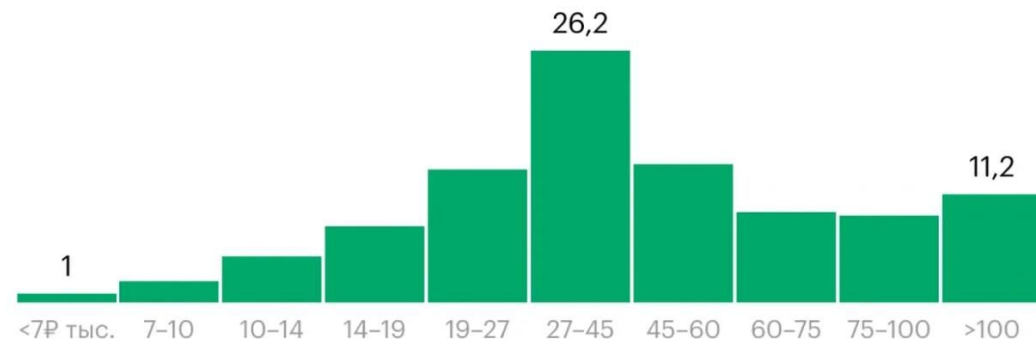


Warmup

Что “не так” с гистограммой?

Распределение населения России по величине среднедушевых доходов

Данные за 2023 год, %



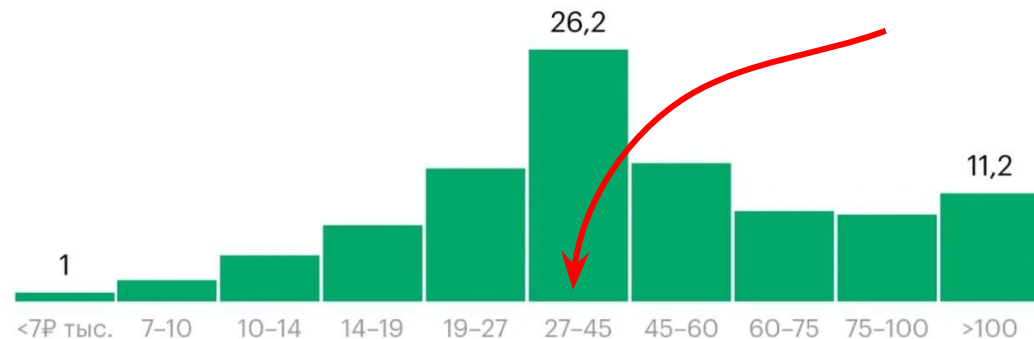
Нижняя граница интервалов не включена

Warmup

Что “не так” с гистограммой?

Распределение населения России по величине среднедушевых доходов

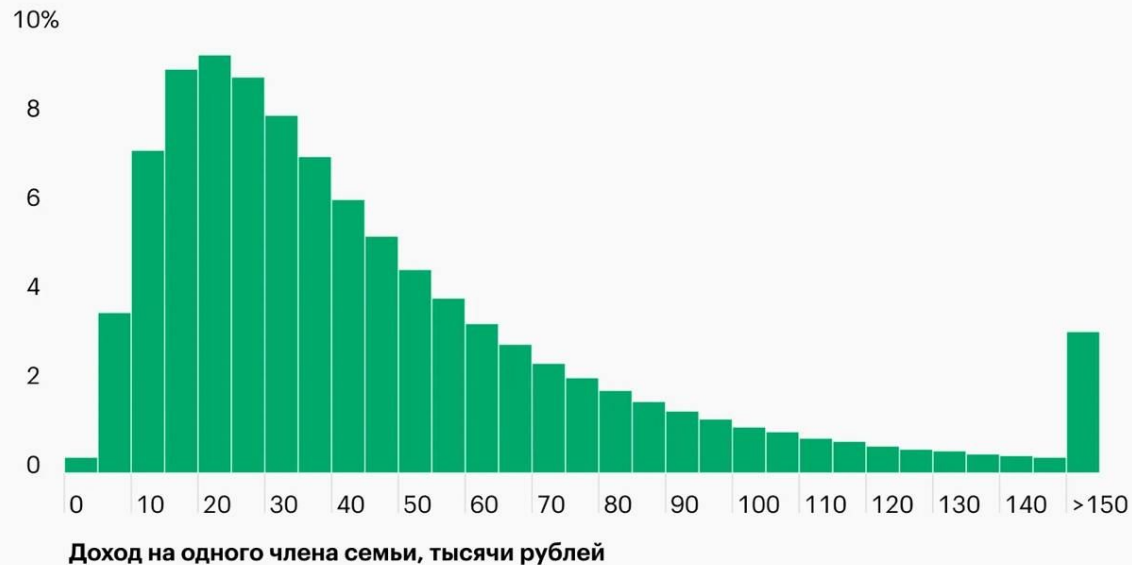
Данные за 2023 год, %



Нижняя граница интервалов не включена

Warmup

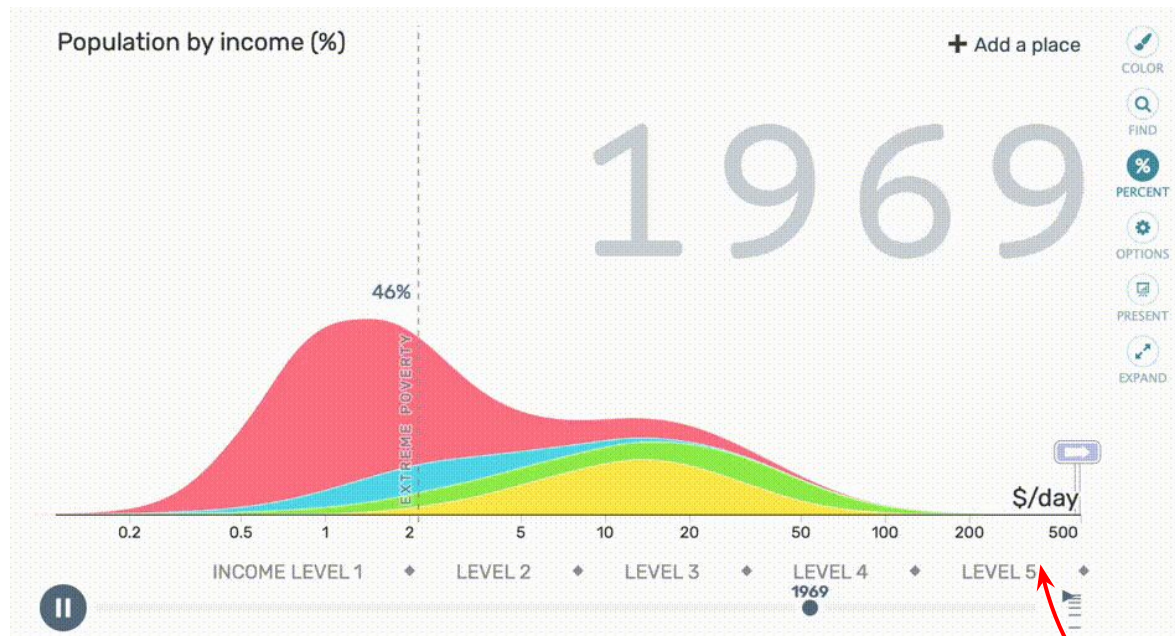
Распределение населения России по величине среднедушевых доходов в 2024 году



Источник:

<https://journal.tinkoff.ru/stat-manipulation-logical/>

Warmup



Источник:

<https://www.gapminder.org/fw/income-mountains/>

Половина людей в мире живут на сумму менее 8 долларов в день

А тут тоже?

Нюанс 1: Паритет покупательской способности

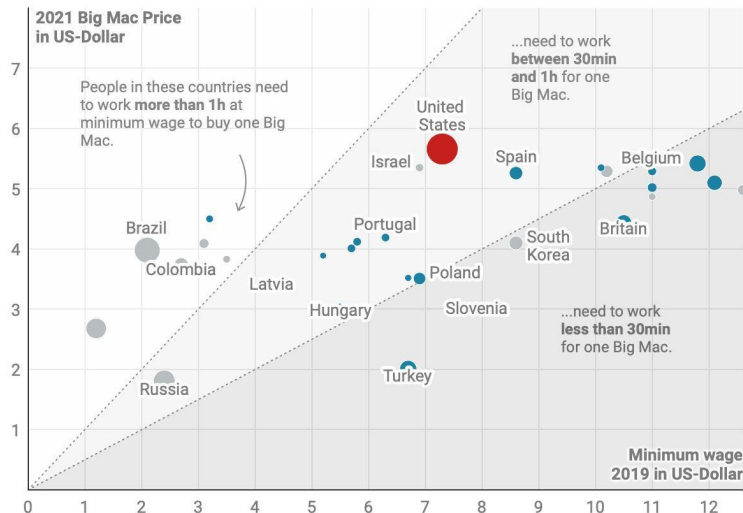
How long do people need to work at minimum wage to buy one Big Mac?

2021 Big Mac price in US-Dollar according to the [Economist Big Mac Index](#) and 2019 minimum wage in US-Dollar in selected countries. The bigger a country circle, the bigger the population. **European countries are highlighted in blue.**

Источник:

https://www.datawrapper.de/_/UsFlq/

https://ru.wikipedia.org/wiki/Индекс_бигмака



Based on a chart by [RJ Andrews](#).

Source: [The Economist \(Big Mac prices\)](#) and [OECD \(minimum wages\)](#) • [Get the data](#) • Created with [Datawrapper](#)

Где мы?

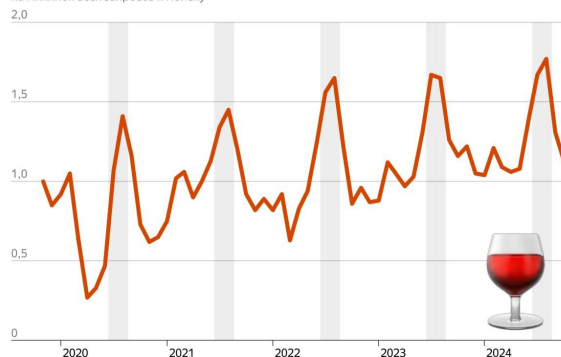
- **Dollar Street** – сайт где можно посмотреть как живут разные семьи в разных странах с определенным уровнем дохода,
- ... по многим параметрам – от домов и машин, до косметики, зубных щеток, зубов, состояния ванных комнат и тп.
- Проект делают уже 15 лет,
- 30 000 фотографий
- 264 семьи со всего мира
- Ссылка: <https://www.gapminder.org>

Рекомендации

- http://t.me/data_csv – канал про дата-журналистику
- <http://t.me/YaResearches> – канал от дата-журналистов Яндекса

Количество запросов про дегустацию вина

на миллион всех запросов к Поиску

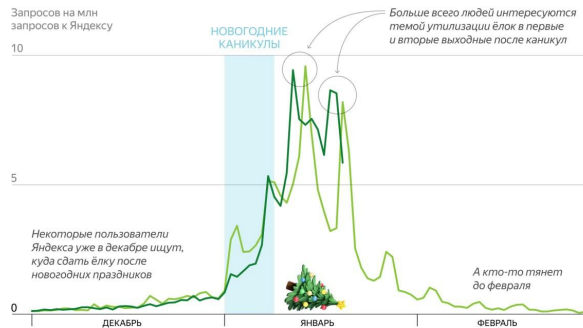


На графике выделены июль-август каждого года

ПО ДАННЫМ ПОИСКА ЯНДЕКСА, НОЯБРЬ 2019 — ОКТЯБРЬ 2024

Как меняется интерес к утилизации ёлок

— Нынешняя зима (2024–2025) — Прошлая зима (2023–2024)



Изучали запросы со словами сдать ёлку, приём ёлок, утилизация ёлок, переработка ёлок, выбросить ёлку.

ПО ДАННЫМ ПОИСКА ЯНДЕКСА

Рекомендации

- http://t.me/data_csv – кан
- <http://t.me/YaResearches>

Интерес к модным игрушкам в поиске Яндекса

Доля от всех запросов, %

0,08



Исследования Яндекса. По данным Поиска



Если запись не включена,
КРИЧИТЕ 😱!

План

План



1. Ловушки визуализации и анализа
2. Основные модули для визуализации
 - Pandas **built-in**
 - Plotly
 - Matplotlib
 - ...

План



1. Ловушки визуализации и анализа
2. Основные модули для визуализации
 - Pandas **built-in**
 - Plotly
 - Matplotlib

Зачем?



- **EDA** = Exploratory Data Analysis = Разведочный анализ данных
- Мониторинг качества
- Помощь в принятии решения

Зачем?



- EDA = Exploratory Data Aalysis = Разведочный анализ данных
- Мониторинг качества
- Помощь в принятии решения

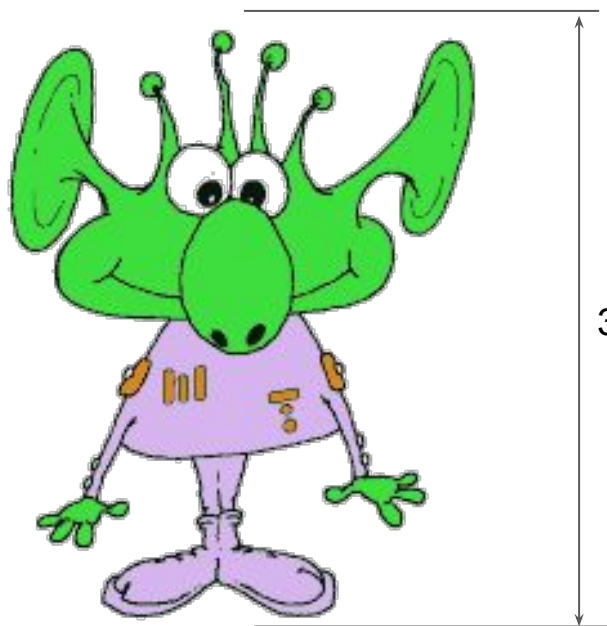
Постановка задачи

Этапы обучения



1. Освоение
- 2. Осмысление**
3. Практика

Мируквите Борука, 107 лет



Министерство здравоохранения СССР
(наименование учреждения)

Код формы по ОКУД _____
Код учреждения по ОКПО _____
Медицинская документация
Форма № 216/у
Утверждена Минздравом СССР
04.10.80 г. № 1030

Лаборатория _____

АНАЛИЗ МОКРОТЫ № _____

« _____ » _____ 20 ____ г. _____ час. _____ мин.

Фамилия, и., о. Мируквите Борука Возраст 107

Учреждение _____ отделение _____ медицинская карта № _____

Палата _____ участок _____

Соленость слезы 93 Запах _____

Цвет 51 Характер вредный


Температура не обнаружены

Пупырки _____

Все измерения производятся в условных единицах измерения, одинаковых для соответствующих значений

Кейс: Решение должно быть принято



 Drive

+ New

Home

My Drive

Computers

Shared with me

Recent

Starred

Search in Drive

≡



... > 02-1: Set description > demo ▾ 👤

Type ▾

People ▾

Modified ▾

Source ▾

Name ▾	Owner
 fars_diseases 👤	 me



Кейс: Может бы так?

1. Научиться делать **сводные таблицы**
2. Придумать “хороший” **показатель**

Диагноз	Хороший показатель
Здоровый	1
Оптимизм 1 степени	7
Оптимизм 2 степени	-5

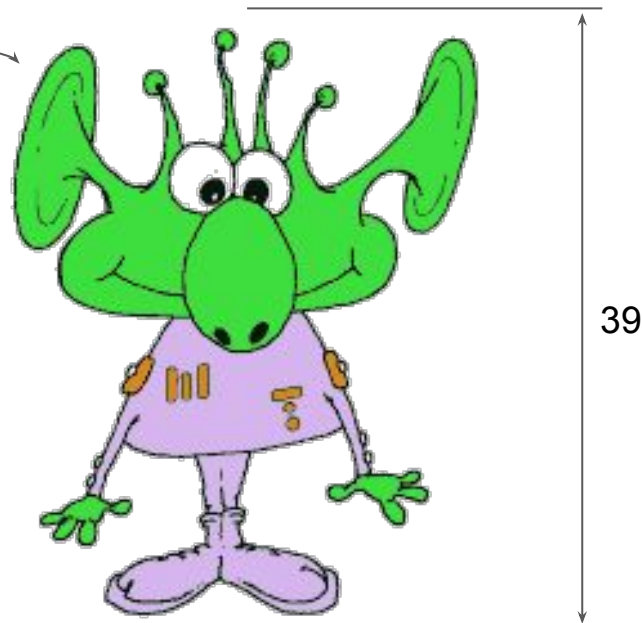


Кейс: Лучше диапазон



1. Научиться делать **сводные таблицы**
2. Придумать “хороший” **показатель**

Диагноз	Хороший показатель
Здоровый	1 – 3
Оптимизм 1 степени	7 – ...
Оптимизм 2 степени	-... – -5



Кейс: Пример



Клинический анализ крови

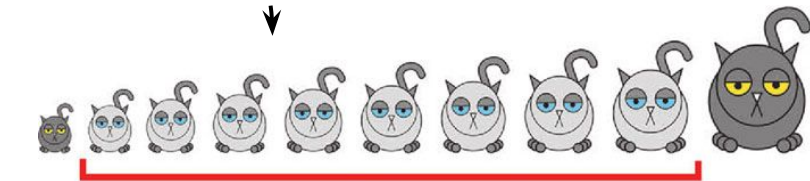
Исследование	Результат	Единицы	Референсные значения
Гематокрит	38.6*	%	39.0 - 49.0
Гемоглобин	13.8	г/дл	13.2 - 17.3
Эритроциты	4.60	млн/мкл	4.30 - 5.70
MCV (ср. объем эритроц.)	83.9	фл	80.0 - 99.0
RDW (шир. распредел. эритроц.)	12.1	%	11.6 - 14.8
MCH (ср. содерж. Hb в эритроц.)	30.0	пг	27.0 - 34.0
MCHC (ср. конц. Hb в эритроц.)	35.8	г/дл	32.0 - 37.0
Тромбоциты	159	тыс/мкл	150 - 400
Лейкоциты	4.69	тыс/мкл	4.50 - 11.00
Нейтрофилы (общ. число), %	42.5*	%	48.0 - 78.0



Кейс: Цель

1. Научиться делать **сводные таблицы**
2. Придумать “хороший” **показатель**

Диагноз	Среднее давление
Здоровый	1
Оптимизм 1 степени	70000
Оптимизм 2 степени	-50000



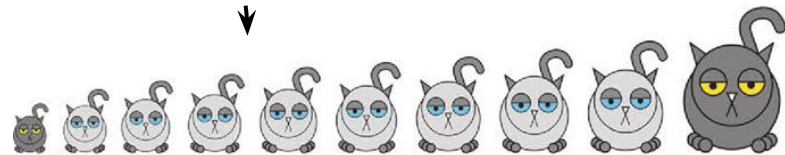
Котики для усеченного среднего



Кейс: Цель

1. Научиться делать **сводные таблицы**
2. Понять, как выбирать **интервал**

Диагноз	Среднее давление
Здоровый	1 – 3
Оптимизм 1 степени	7 – ...
Оптимизм 2 степени	-... – -5



Котики для усеченного среднего

Кейс: Как получить доступ?



The screenshot shows the Google Sheets interface for a file named 'fars_diseases'. The 'File' menu is open, displaying options: Import, Make a copy, Share, Email, Download, Rename, and Move. The 'Share' option is selected, and its sub-menu is visible, showing 'Share with others' and 'Publish to web'. The 'Publish to web' option is highlighted. The background shows a spreadsheet with columns labeled 'A' through 'G' and rows numbered 1 through 14. The data in the spreadsheet includes numerical values and the word 'TRUE'.

	A	B	C	D	E	F	G
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							

Группировки

Фамилия	Район проживания	Соленость слезы	Количество ушей	Диагноз
Бихивеин	Адок	137	7	Оптимизм 2 степени
Бихошюев	Адок	72	7	Оптимизм 2 степени
Бидетаин	Хоккинг-стрит	88	5	Повышенная пупырчатость
Бицевюев	Хоккинг-стрит	90	5	Здоров
Бицедиова	Хоккинг-стрит	10	4	Оптимизм 1 степени
Бацегаин	Площадь Гагарина	20	4	Трихотомия
Бачытиин	Площадь Гагарина	60	4	Повышенная пупырчатость
Бикэнаова	Площадь Гагарина	67	0	Оптимизм 1 степени

```
data[data["Количество ушей"]==4]["Соленость слезы"].mean()
```

Фамилия	Район проживания	Соленость слезы	Количество ушей	Диагноз
Бихивеин	Адок	137	7	Оптимизм 2 степени
Бихошюев	Адок	72	7	Оптимизм 2 степени
Бидетаин	Хоккинг-стрит	88	5	Повышенная пупырчатость
Бицевюев	Хоккинг-стрит	90	5	Здоров
Бигедиова	Хоккинг-стрит	10	4	Оптимизм 1 степени
Бацегаин	Площадь Гагарина	20	4	Трихотомия
Бачытиин	Площадь Гагарина	60	4	Повышенная пупырчатость
Бикэнаова	Площадь Гагарина	67	0	Оптимизм 1 степени

`data[data["Количество ушей"]==4][["Соленость слезы"]].max()`

Фильтр
Отбор столбцов
Агрегация

Фамилия	Район проживания	Соленость слезы	Количество ушей	Диагноз
Бихивеин	Адок	137	7	Оптимизм 2 степени
Бихошюев	Адок	72	7	Оптимизм 2 степени
Бидетаин	Хоккинг-стрит	88	5	Повышенная пупырчатость
Бицевюев	Хоккинг-стрит	90	5	Здоров
Бицедиова	Хоккинг-стрит	10	4	Оптимизм 1 степени
Бацегаин	Площадь Гагарина	20	4	Трихотомия
Бачытиин	Площадь Гагарина	60	4	Повышенная пупырчатость
Бикэнаова	Площадь Гагарина	67	0	Оптимизм 1 степени

`data` [`data` ["Количество ушей"] == 4] ["Соленость слезы"] .max()

Фильтр
Отбор столбцов
Агрегация

Фамилия	Район проживания	Соленость слезы	Количество ушей	Диагноз
Бихивеин	Адок	137	7	Оптимизм 2 степени
Бихошюев	Адок	72	7	Оптимизм 2 степени
Бидетаин	Хоккинг-стрит	88	5	Повышенная пупырчатость
Бицевюев	Хоккинг-стрит	90	5	Здоров
Бицедиова	Хоккинг-стрит	10	4	Оптимизм 1 степени
Бацегаин	Площадь Гагарина	20	4	Трихотомия
Бачытиин	Площадь Гагарина	60	4	Повышенная пупырчатость
Бикэнаова	Площадь Гагарина	67	0	Оптимизм 1 степени

`data[data["Количество ушей"]==4]["Соленость слезы"].max()`

{ } { } { }

Фильтр Отбор столбцов Агрегация

Фамилия	Район проживания	Соленость слезы	Количество ушей	Диагноз
Бихивеин	Адок	137	7	Оптимизм 2 степени
Бихошюев	Адок	72	7	Оптимизм 2 степени
Бидетаин	Хоккинг-стрит	88	5	Повышенная пупырчатость
Бицевюев	Хоккинг-стрит	90	5	Здоров
Бицедиова	Хоккинг-стрит	10	4	Оптимизм 1 степени
Бацегаин	Площадь Гагарина	20	4	Трихотомия
Бачытиин	Площадь Гагарина	60	4	Повышенная пупырчатость
Бикэнаова	Площадь Гагарина	67	0	Оптимизм 1 степени

`data[data["Количество ушей"]==4]["Соленость слезы"] .max()`

Фильтр
Отбор столбцов
Агрегация

Фамилия	Район проживания	Соленость слезы	Количество ушей	Диагноз
Бихивеин	Адок	137	7	Оптимизм 2 степени
Бихошюев	Адок	72	7	Оптимизм 2 степени
Бидетаин	Хоккинг-стрит	88	5	Повышенная пупырчатость
Бицевюев	Хоккинг-стрит	90	5	Здоров
Бицедиова	Хоккинг-стрит	10	4	Оптимизм 1 степени
Бацегаин	Площадь Гагарина	20	4	Трихотомия
Бачытиин	Площадь Гагарина	60	4	Повышенная пупырчатость
Бикэнаова	Площадь Гагарина	67	0	Оптимизм 1 степени

`data[data["Количество ушей"]==4]["Соленость слезы"] .min()`

{ } { } { }

Фильтр Отбор столбцов Агрегация

Фамилия	Район проживания	Соленость слезы	Количество ушей	Диагноз
Бихивеин	Адок	137	7	Оптимизм 2 степени
Бихошюев	Адок	72	7	Оптимизм 2 степени
Бидетаин	Хоккинг-стрит	88	5	Повышенная пупырчатость
Бицевюев	Хоккинг-стрит	90	5	Здоров
Бицедиова	Хоккинг-стрит	10	4	Оптимизм 1 степени
Бацегаин	Площадь Гагарина	20	4	Трихотомия
Бачытиин	Площадь Гагарина	60	4	Повышенная пупырчатость
Бикэнаова	Площадь Гагарина	67	0	Оптимизм 1 степени

```
data[data["Количество ушей"]==4][["Соленость слезы"]].count()
```

Фильтр

Отбор столбцов

Агрегация

Фамилия	Район проживания	Соленость слезы	Количество ушей	Диагноз
Бихивеин	Адок	137	7	Оптимизм 2 степени
Бихошюев	Адок	72	7	Оптимизм 2 степени
Бидетаин	Хоккинг-стрит	88	5	Повышенная пупырчатость
Бицевюев	Хоккинг-стрит	90	5	Здоров
Бицедиова	Хоккинг-стрит	10	4	Оптимизм 1 степени
Бацегаин	Площадь Гагарина	20	4	Трихотомия
Бачытиин	Площадь Гагарина	60	4	Повышенная пупырчатость
Бикэнаова	Площадь Гагарина	67	0	Оптимизм 1 степени

```
data[data["Количество ушей"]==4][["Соленость слезы"]].mean()
```



Фильтр



Отбор столбцов



Агрегация

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	max (Соленость слезы)
Оптимизм 2 степени	137
Пупырчатость	90
Здоров	67

Хотим так ...

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	max (Соленость слезы)
Оптимизм 2 степени	137

А, если только для “Оптимизма 2 степени”?

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	max (Соленость слезы)
Оптимизм 2 степени	137

А, если только для “Оптимизма 2 степени”?

```
data[data["Диагноз"] == "Оптимизм 2 степени"]
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	max (Соленость слезы)
Оптимизм 2 степени	137

А, если только для “Оптимизма 2 степени”?

```
data[data["Диагноз"] == "Оптимизм 2 степени"] ["Соленость"]
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	max (Соленость слезы)
Оптимизм 2 степени	137

А, если только для “Оптимизма 2 степени”?

```
data[data["Диагноз"] == "Оптимизм 2 степени"]["Соленость"].max()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	max (Соленость слезы)
Оптимизм 2 степени	137
Пупырчатость	90
Здоров	67

Теперь для всех!

```
data[data["Диагноз"] == "Оптимизм 2 степени"]["Соленость"].max()
```


Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	max (Соленость слезы)
Оптимизм 2 степени	137
Пупырчатость	90
Здоров	67

Теперь для всех!

```
data[data["Диагноз"] == "Оптимизм 2 степени"]["Соленость"].max()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	max (Соленость слезы)
Оптимизм 2 степени	137
Пупырчатость	90
Здоров	67

Теперь для всех!

```
data.groupby(by="Диагноз")["Соленость"].max()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
Оптимизм 2 степени	<u>137</u>
Пупырчатость	<u>90</u>
Здоров	<u>67</u>

```
data.groupby(by=' Диагноз ' ) [ ' Соленость слезы ' ] .max ( )
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
Оптимизм 2 степени	<u>137</u>
Пупырчатость	<u>90</u>
Здоров	<u>67</u>

`data.groupby(by=' Диагноз ') [' Соленость слезы '].max()`

`data[data [' Диагноз '] == 'Пупырчатость'] [' Соленость слезы '].max()`

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
Оптимизм 2 степени	
Пупырчатость	
Здоров	

```
data.groupby(by=' Диагноз ' ) [ ' Соленость слезы ' ] .min()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
Оптимизм 2 степени	<u>88</u>
Пупырчатость	<u>90</u>
Здоров	<u>10</u>

```
data.groupby(by=' Диагноз ' ) [ ' Соленость слезы ' ] .min()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
Оптимизм 2 степени	
Пупырчатость	
Здоров	

```
data.groupby(by=' Диагноз ' ) [ ' Количество ушей ' ] .min()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
Оптимизм 2 степени	5
Пупырчатость	5
Здоров	0

```
data.groupby(by=' Диагноз ' ) [ ' Количество ушей ' ] .min()
```


Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	

```
data.groupby(by='Количество ушей')['Диагноз'].min()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

7	
5	
4	
0	

```
data.groupby(by='Количество ушей')['Диагноз'].min()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
0	Здоров
4	Здоров
5	Оптимизим
7	Оптимизм

```
data.groupby(by='Количество ушей')['Диагноз'].min()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
7	
5	
4	
0	

```
data.groupby(by='Количество ушей') ['Диагноз'].count()
```

Фамилия	Соленость слезы	Количество ушей	Диагноз
Бихивеин	137	7	Оптимизм 2 степени
Бихошюев	72	7	Оптимизм 2 степени
Бидетаин	88	5	Оптимизм 2 степени
Бицевюев	90	5	Пупырчатость
Бицедиова	10	4	Здоров
Бацегаин	20	4	Здоров
Бачытиин	60	4	Здоров
Бикэнаова	67	0	Здоров

Диагноз	
7	2
5	2
4	3
0	1

```
data.groupby(by='Количество ушей') ['Диагноз'].count()
```

Что получилось?



Кейс: Итог

```
data.groupby(by='Диагноз')['Соленость слезы'].mean()
```



Соленость слезы

Диагноз

Грибб	103.673469
Дихотомия	101.566667
Здоров	100.287129
Оптимизм 1 степени	104.160000
Оптимизм 2 степени	102.820000
Повышенная пупырчатость	99.700000
Трихотомия	99.510204

dtype: float64



Кейс: Визуализация

```
data.groupby(by='Диагноз')['Соленость слезы'].mean()
```



Соленость слезы

Диагноз

Грибб	103.673469
Дихотомия	101.566667
Здоров	100.287129
Оптимизм 1 степени	104.160000
Оптимизм 2 степени	102.820000
Повышенная пупырчатость	99.700000
Трихотомия	99.510204

dtype: float64

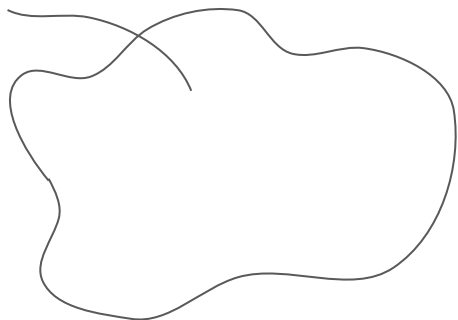
Характеристика множества

Характеристика множества

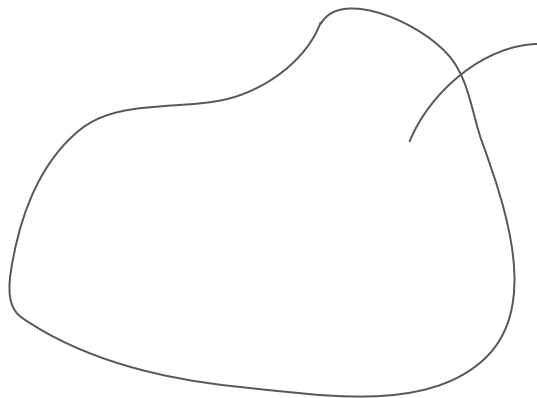


- Кто наши клиенты?
- В каком городе открывать филиал?
- Какая аудитория более лояльная?

42



100 500



Среднее арифметическое



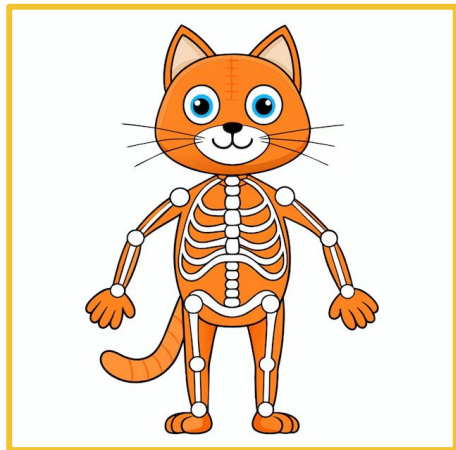
Среднее арифметическое. Плюсы

1. Общеизвестно
2. Легко вычислить
3. Дифференцируемо

$$\frac{x_1 + x_2 + \dots + x_n}{n}$$



Среднее арифметическое. Минусы



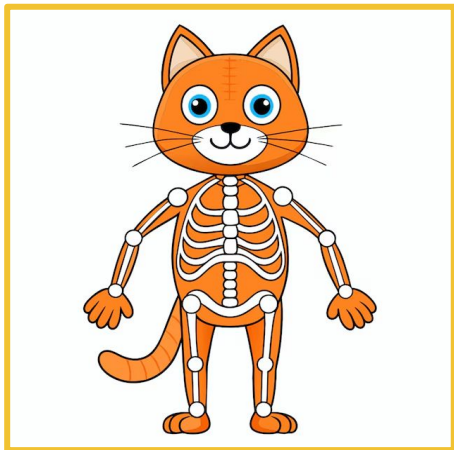
Среднее
число скелетов
в организме кота:

А: Меньше 1

Б: Равно 1

В: Превышает 1

Среднее арифметическое. Минусы



Среднее
число скелетов
в организме кота:

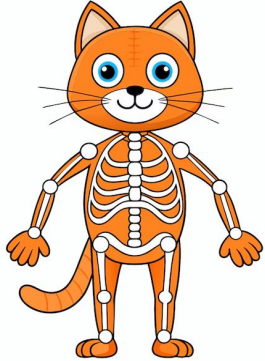
А: Меньше 1

Б: Равно 1

В: Превышает 1



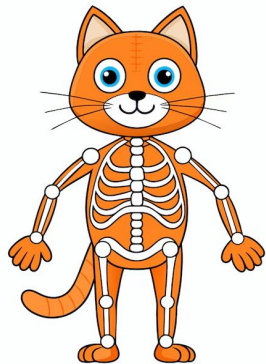
Среднее арифметическое. Минусы



Среднее
число скелетов
в организме кота превышает 1

$$\frac{x_1 + x_2 + \dots + x_n}{n} =$$


Среднее арифметическое. Минусы



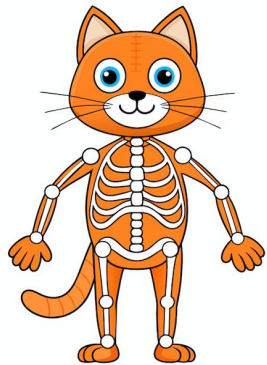
Среднее
число скелетов
в организме кота превышает 1

$$\frac{x_1 + x_2 + \dots + x_n}{n}$$

=


$$5$$

Среднее арифметическое. Минусы



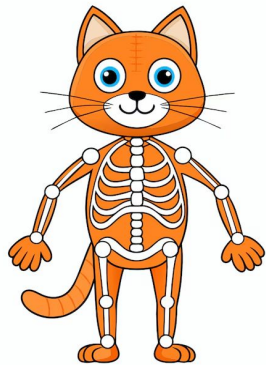
Среднее
число скелетов
в организме кота превышает 1

$$\frac{x_1 + x_2 + \dots + x_n}{n}$$

=

5

Среднее арифметическое. Минусы



Среднее
число скелетов
в организме кота превышает 1

$$\frac{x_1 + x_2 + \dots + x_n}{n}$$

=

$$\frac{\text{12 moon emojis}}{12}$$

Среднее арифметическое. Минусы





1. Может трактоваться неоднозначно

2



Среднее арифметическое. Минусы


1. Может трактоваться неоднозначно
2. Плохой показатель типичности




Предупреждение: у нас есть цензура и предварительный
покиньте сайт. 18+

Анекдоты из России

выходят с 8 ноября 1995 года
составитель - Дима Вернер



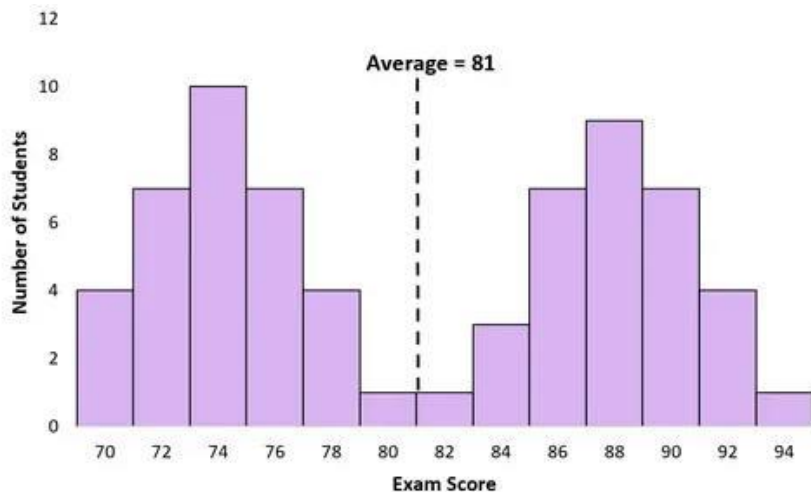
АНЕКДОТ №244368

Согласно статистике, у среднего человека одна грудь и одно 

Среднее арифметическое. Минусы



1. Может трактоваться неоднозначно
2. Плохой показатель типичности



3

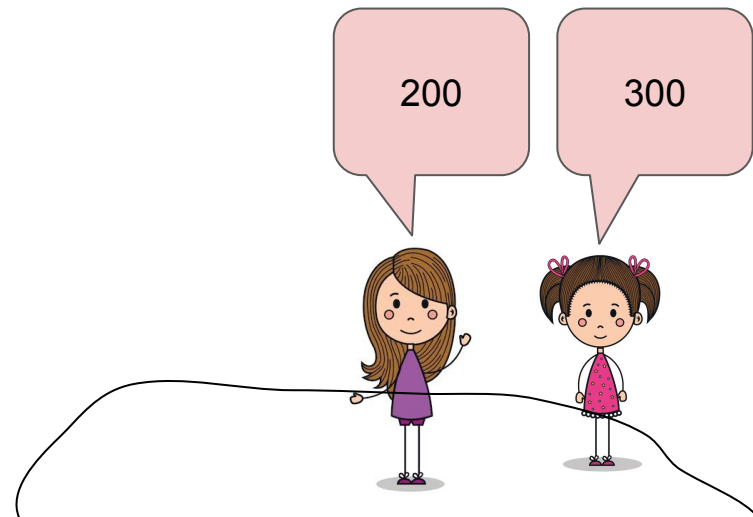
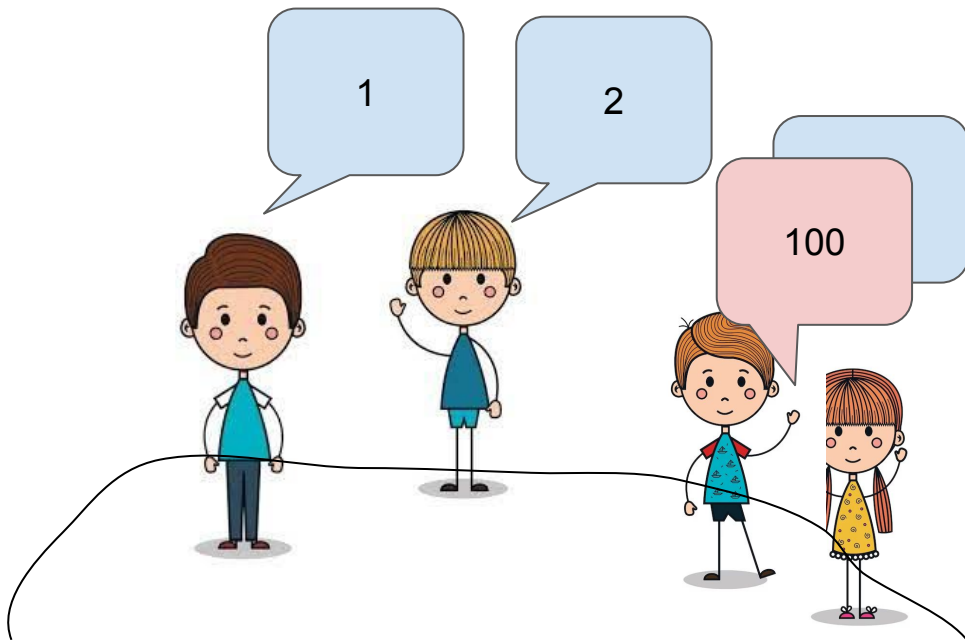
Интуиция



Группа граждан страны А эмигрировала в страну Б.

Мог ли в этих двух странах одновременно вырасти **средний IQ**?

среднее арифметическое



Феномен Уилла Роджерса

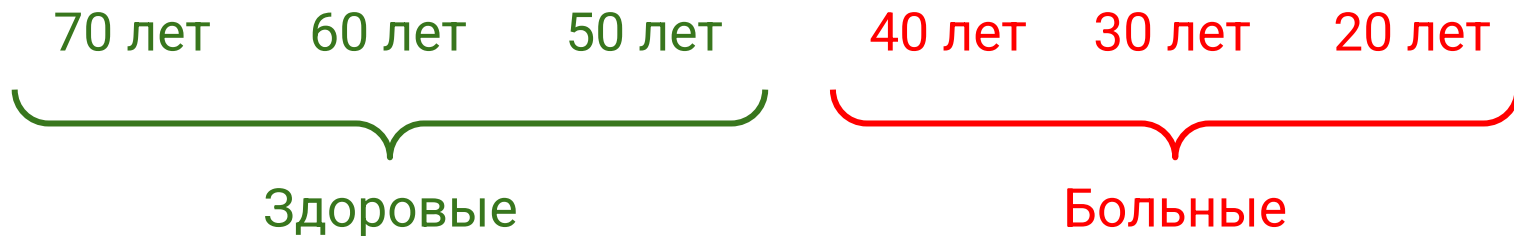


“С 1950 года **больные X** стали жить, в среднем, дольше”

Феномен Уилла Роджерса



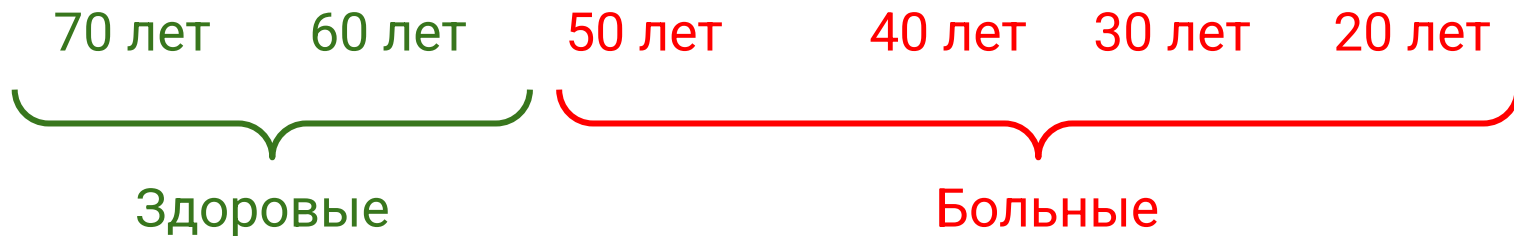
“С 1950 года **больные X** стали жить, в среднем, дольше”



Феномен Уилла Роджерса



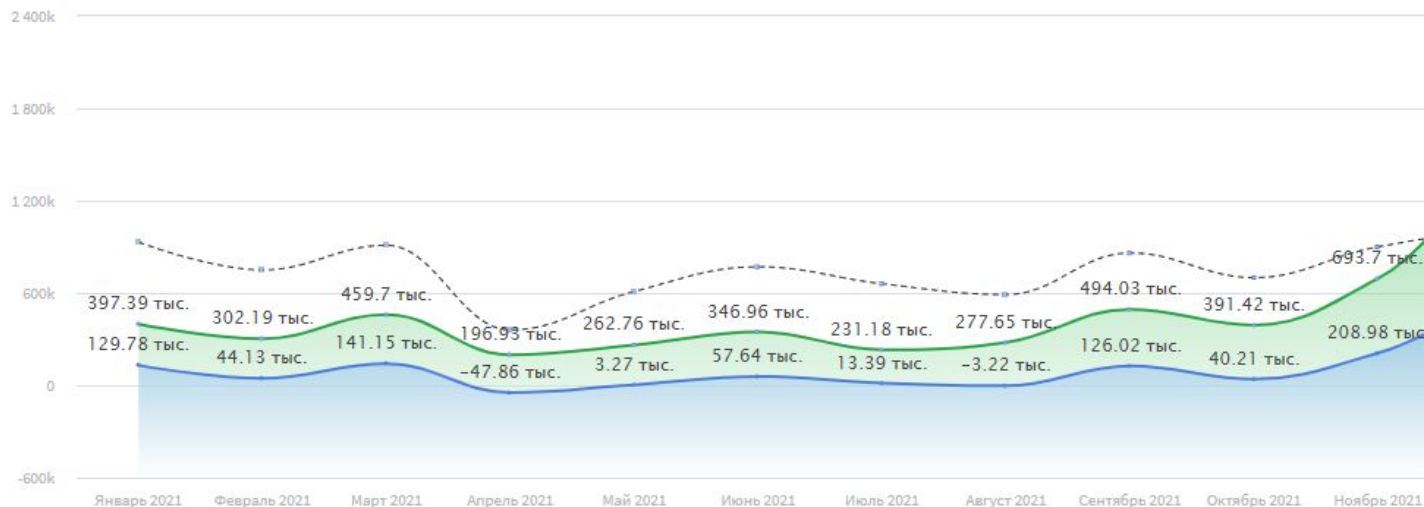
“С 1950 года **больные X** стали жить, в среднем, дольше”



Феномен Уилла Роджерса



По месяцам ▾





Среднее арифметическое. Минусы

1. Может трактоваться неоднозначно
2. Плохой показатель типичности
3. В динамике может вводить в заблуждение

4



Среднее арифметическое. Минусы

1. Может трактоваться неоднозначно
2. Плохой показатель типичности
3. В динамике может вводить в заблуждение
4. Плохо интерпретируем

5

Пример №1

- Журнал “Torbes” хочет построить рейтинг самых высокооплачиваемых профессий.
- Редакторы рассмотрели несколько профессий и узнали зарплаты 5 работников каждой из них.
- Как построить рейтинг? Ваше мнение...

Пример №1

12 8 10 10 10



№1

1 2 3 4 40



№2

Пример №1

- Первое, что приходит в голову: среднее арифметическое.

Пример №1

Посчитайте среднее арифметическое

12 8 10 10 10



№1

1 2 3 4 40



№2

Пример №1

“У нас средняя зарплата 10 тыс. тугриков!”

12 8 10 10 10



№1

1 2 3 4 40



№2

Пример №1

Средняя зарплата одинакова. Но множества совсем разные.

12 8 10 10 10



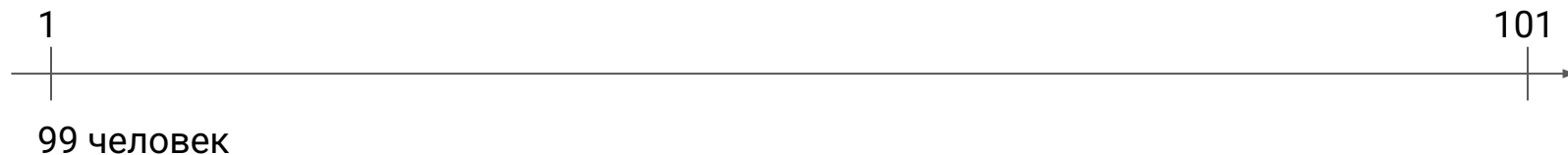
№1

1 2 3 4 40



№2

Еще приколы **среднего арифметического**



Еще приколы среднего арифметического



Еще





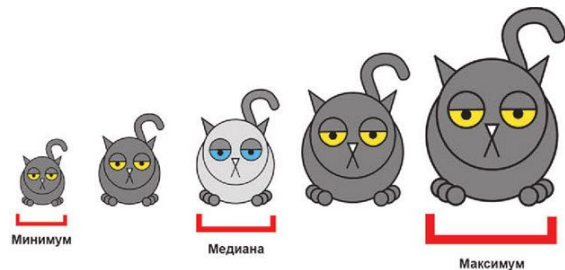
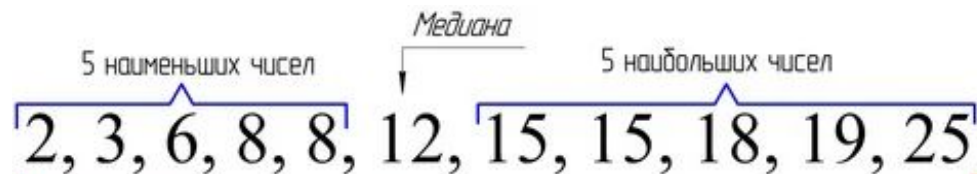
Среднее арифметическое. Минусы

1. Может трактоваться неоднозначно
2. Плохой показатель типичности
3. В динамике может вводить в заблуждение
4. Плохо интерпретируем

Медиана

Медиана конечного множества чисел это – ...

Число, которое находится в середине набора чисел, если его упорядочить по возрастанию.

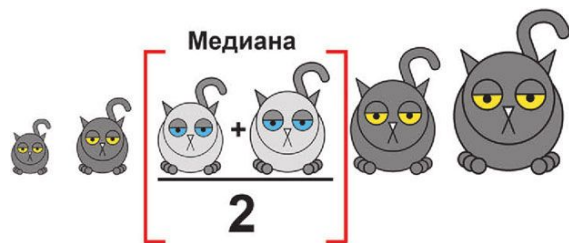
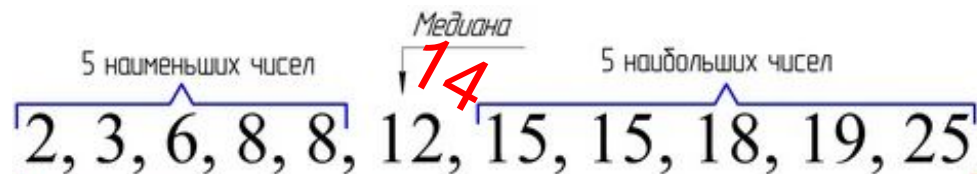


В

Для множества действительных чисел $S = \{a_1, \dots, a_n\}$ такое число $m \in S$, что $|\|\{a_i \in S \mid a_i < m\}\| - \|\{a_i \in S \mid a_i > m\}\|| \leq 1$

Медиана конечного множества чисел это – ...

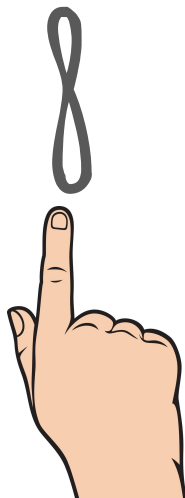
Число, которое находится в середине набора чисел, если его упорядочить по возрастанию.



Для множества действительных чисел $S = \{a_1, \dots, a_n\}$ такое число $m \in S$, что $|\|\{a_i \in S \mid a_i < m\}\| - \|\{a_i \in S \mid a_i > m\}\|| \leq 1$

Укажете на медиану

}



1000

Медиана

8 10 10 10 12



№1

1 2 3 4 40



№2

Медиана

50% слева, 50% справа

8 10 **10** 10 12



№1

1 2 **3** 4 40



№2

Медиана

Если четное количество

8 10 **10** 11 12 12



№1

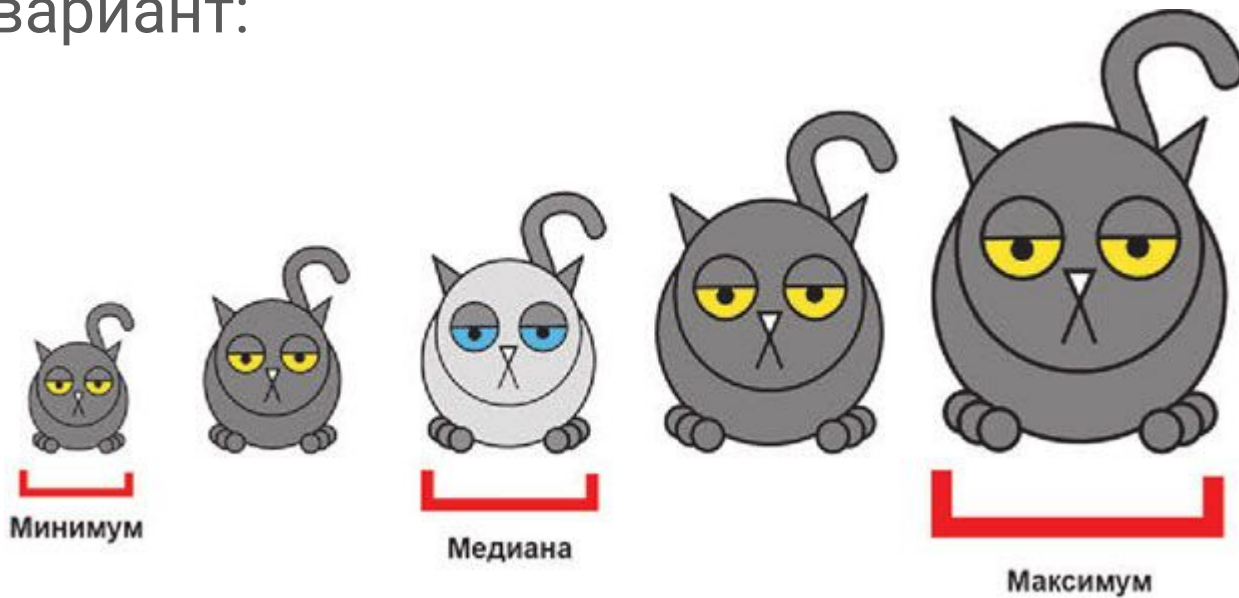
1 2 **3** 4 40



№2

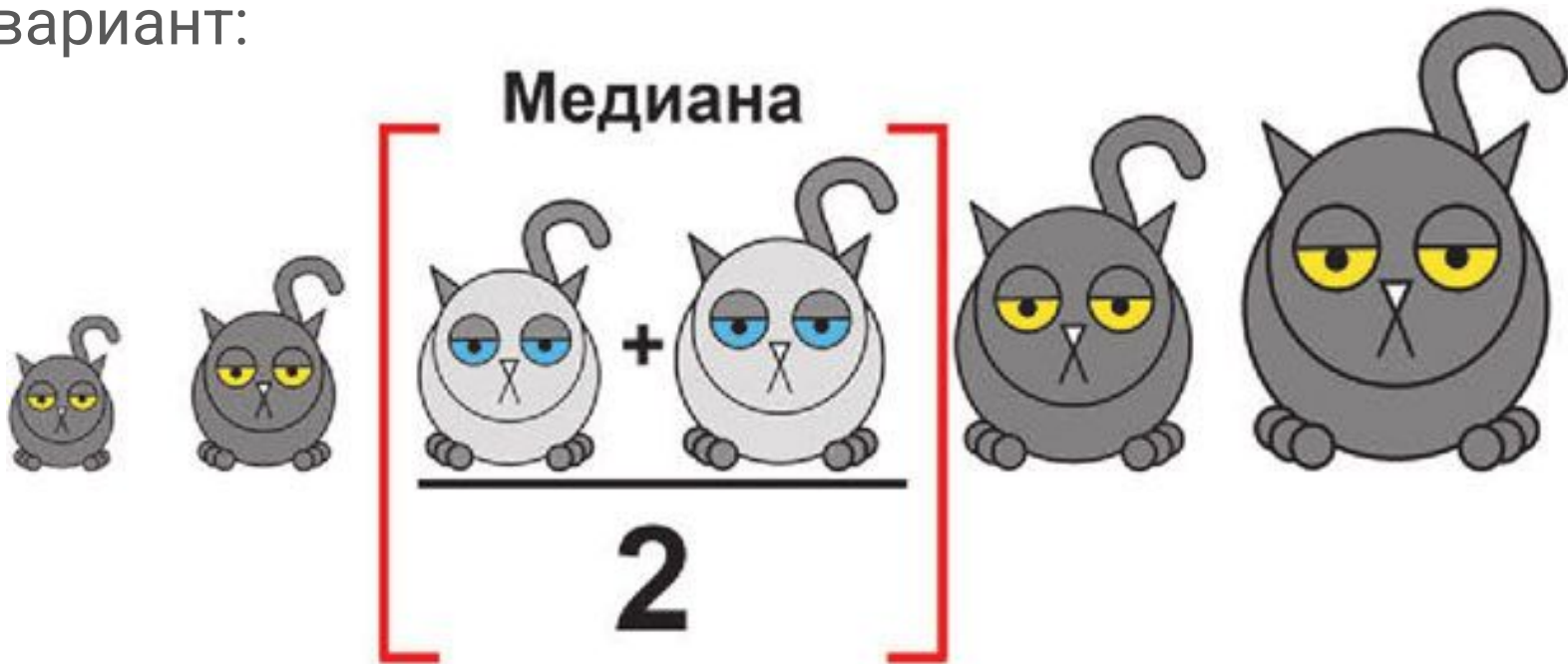
Медиана

Другой вариант:



Медиана

Другой вариант:



Медиана

Свойства

8 10 10 11 12 12



№1

3



№2

Медиана

1. Устойчива к выбросам

8 10 **10** 11 12 12



№1

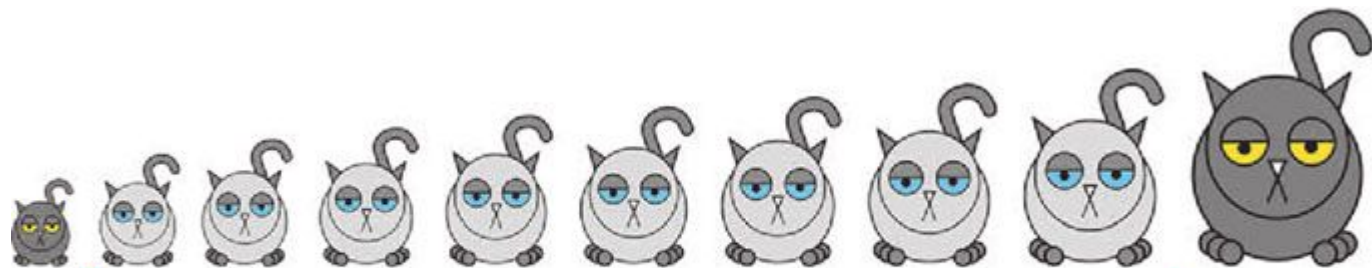
1 2 **3** 4 40 10000



№2



Выброс



Котики для усеченного среднего

Медиана

1. Устойчива к выбросам
2. Указывает на реальный объект

8 10 **10** 11 12 12

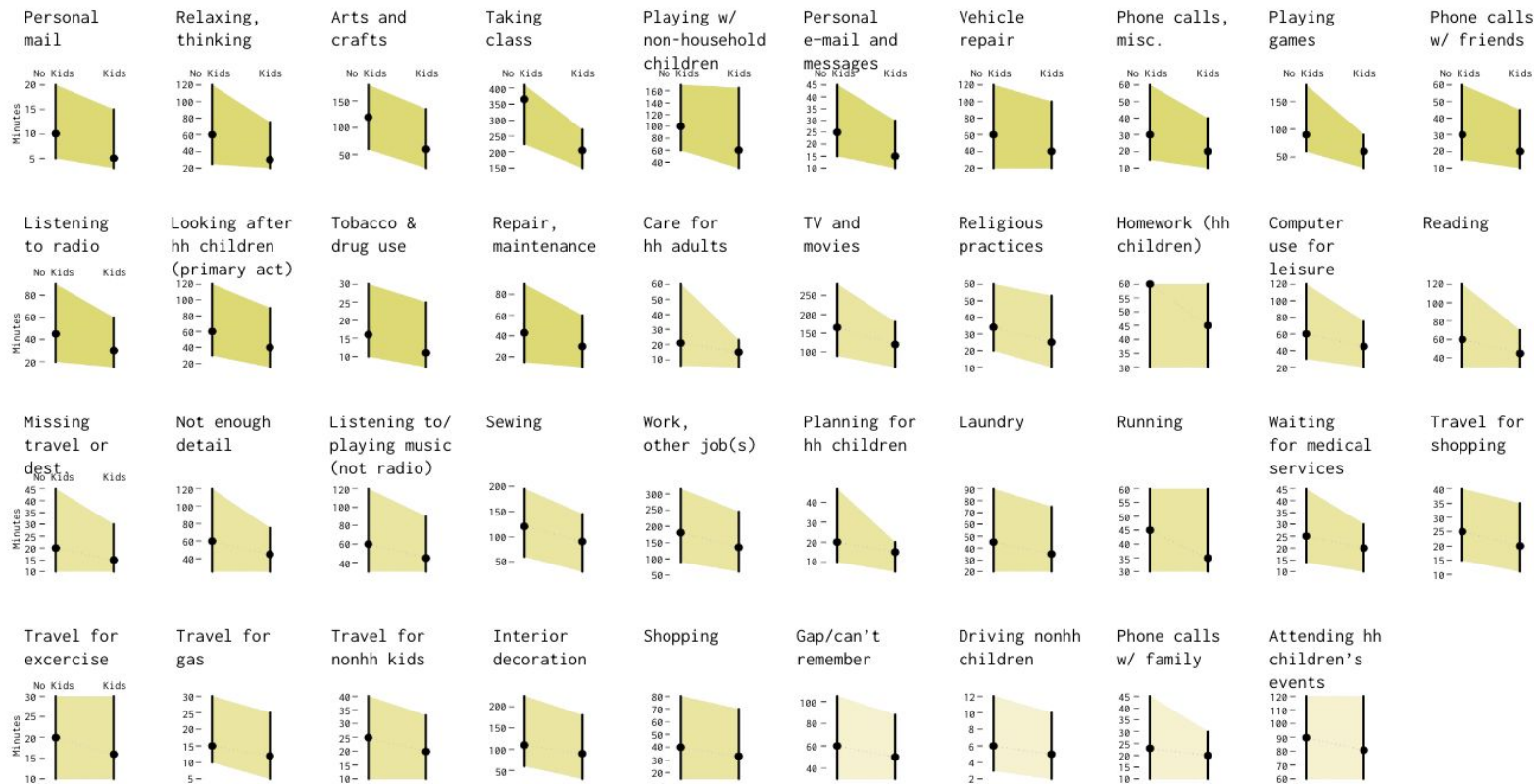


№1

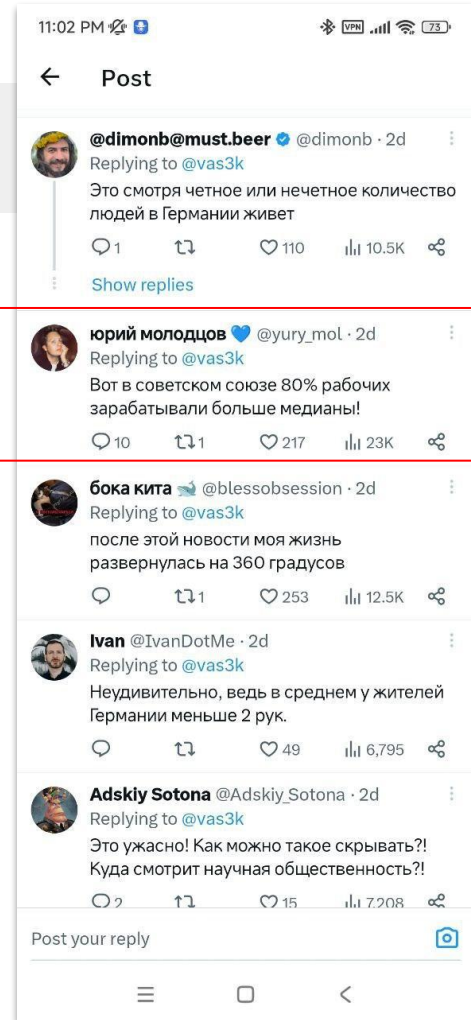
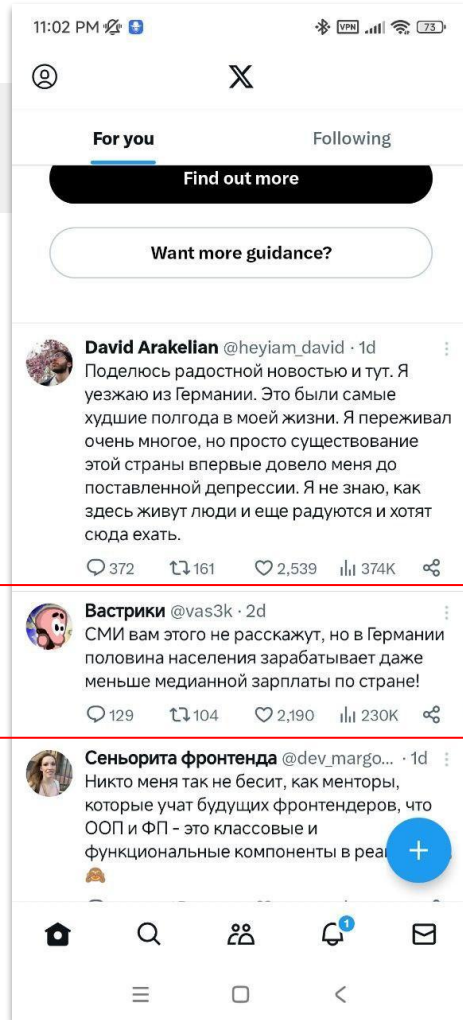
1 2 **3** 4 40 10000



№2



Шуточки на понимание

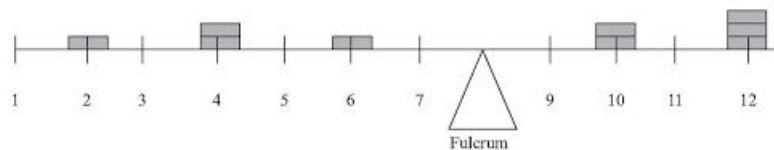
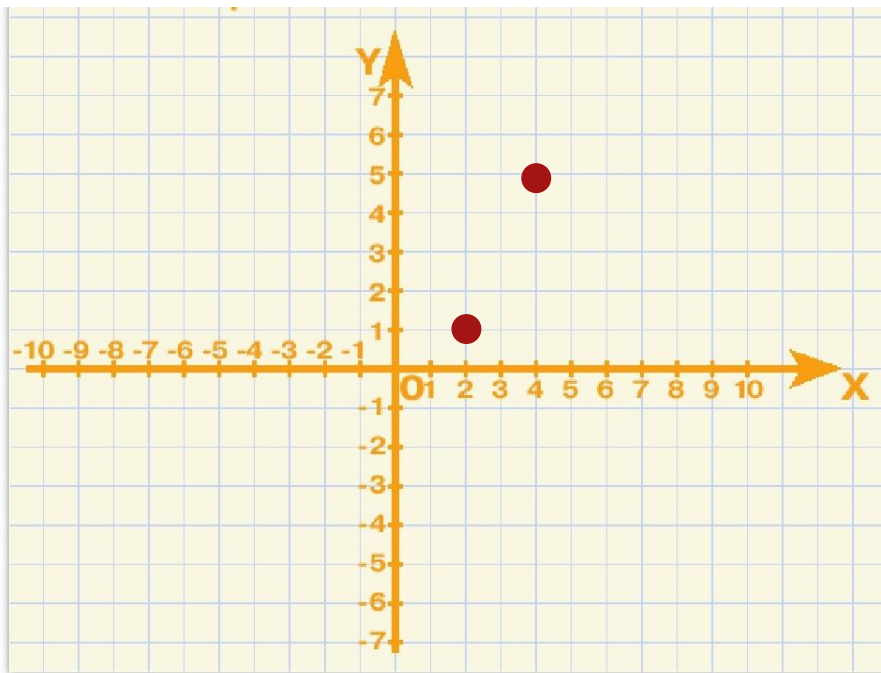


Недостатки

Видимо, одного признака мало ...

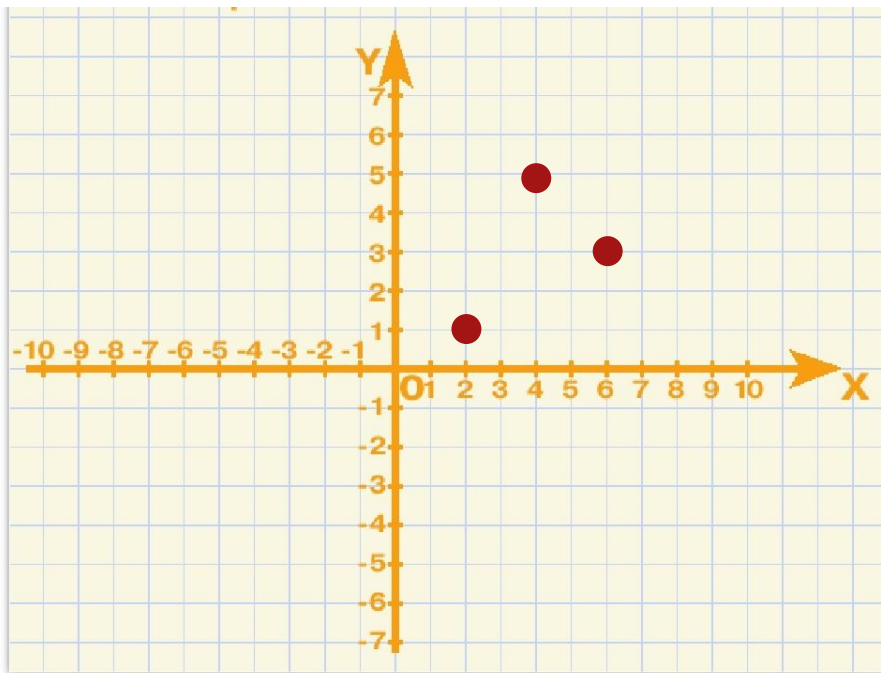
Задача

Среднее арифметическое легко обобщается на \mathbb{R}_2 :



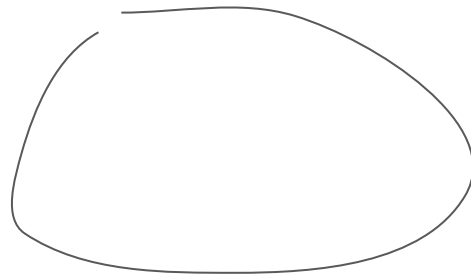
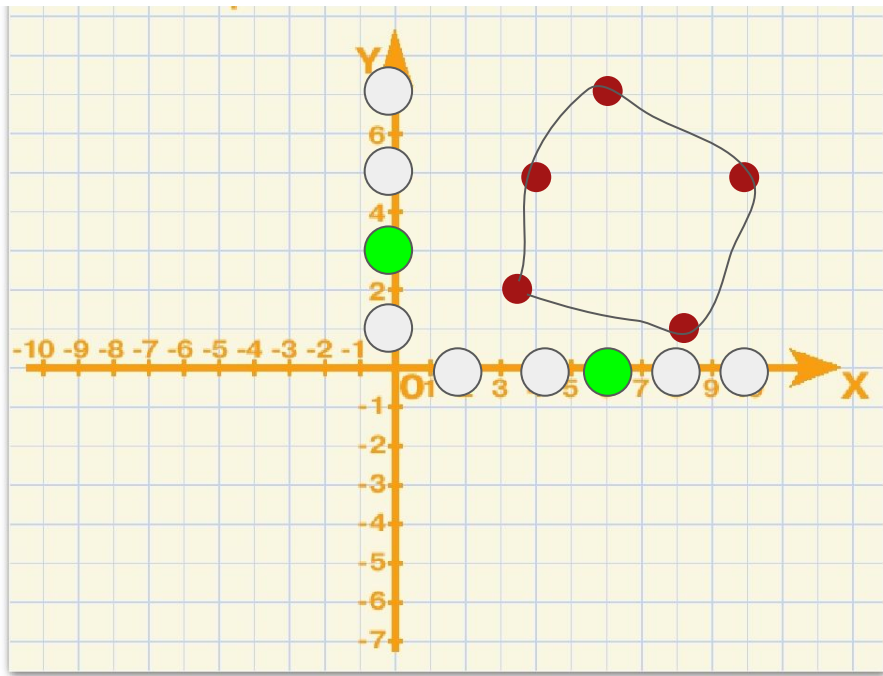
Задача

Среднее арифметическое легко обобщается на \mathbb{R}_2 :



Задача

Как обобщить понятие **медианы** на плоскость (\mathbb{R}_2)?



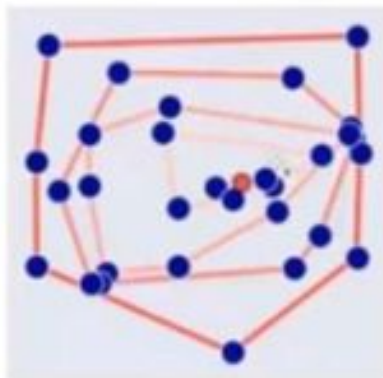
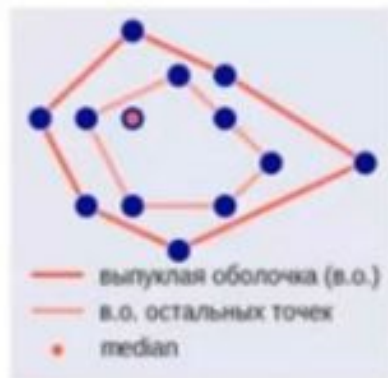
Другие
характеристики

Задача

Как обобщить понятие медианы на плоскость (\mathbb{R}_2)?

Какие важные свойства будем сохранять?

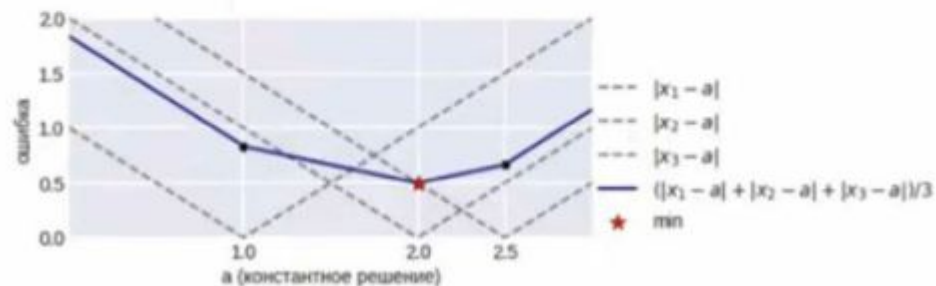
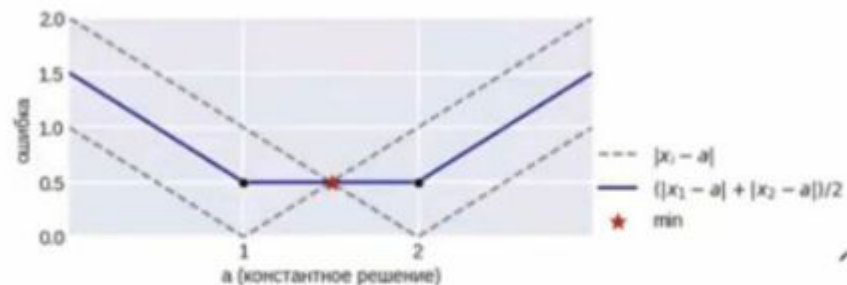
Вариант 1. Обобщить способ построения



**Выход: сделать аналогичный процесс построения,
как в одномерном случае
удаление крайних элементов!**

Вариант 2. Обобщить задачу

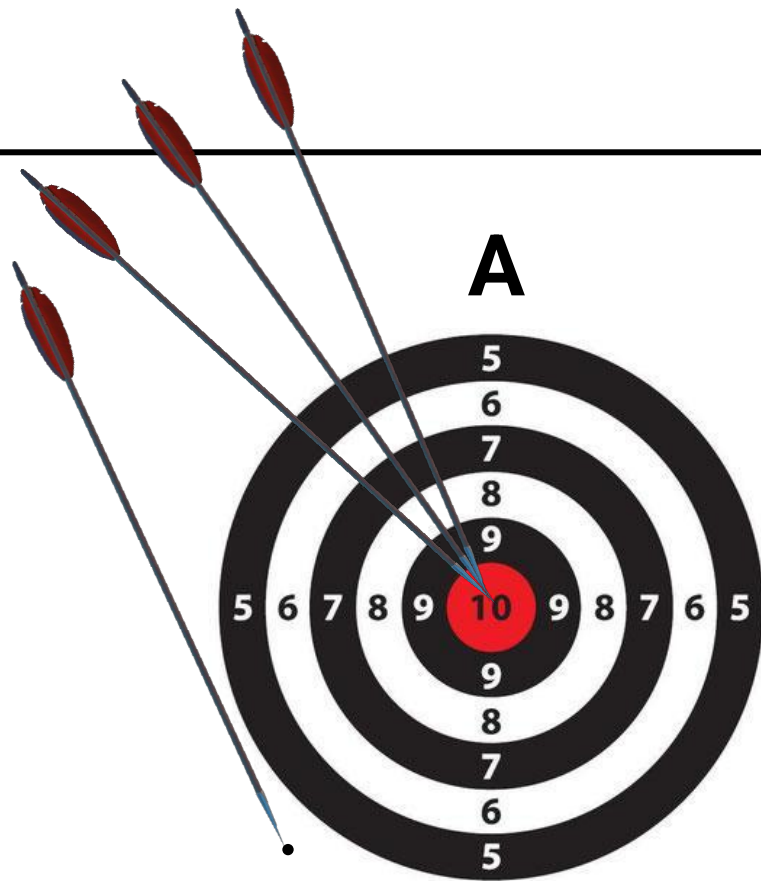
Среднее как решение оптимизационной задачи



© 2010 Pearson Education, Inc.
All rights reserved.
This publication is for the exclusive use of the individual user.
No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or by any information storage or retrieval system, without prior written permission from Pearson Education, Inc.

Мода

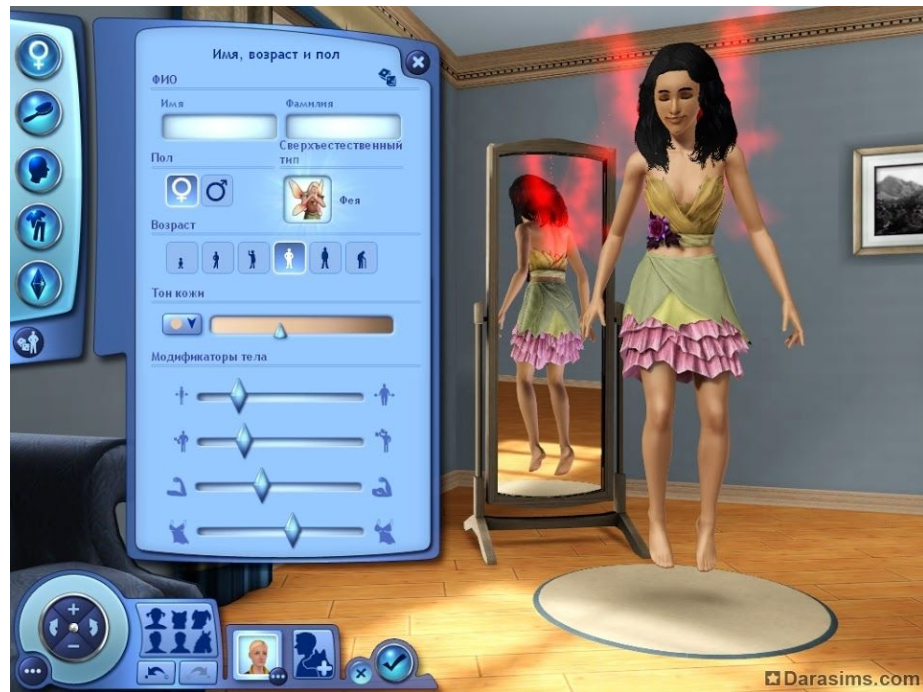
Одно или несколько значений во множестве, которое встречается наиболее часто.



Типичный россиянин

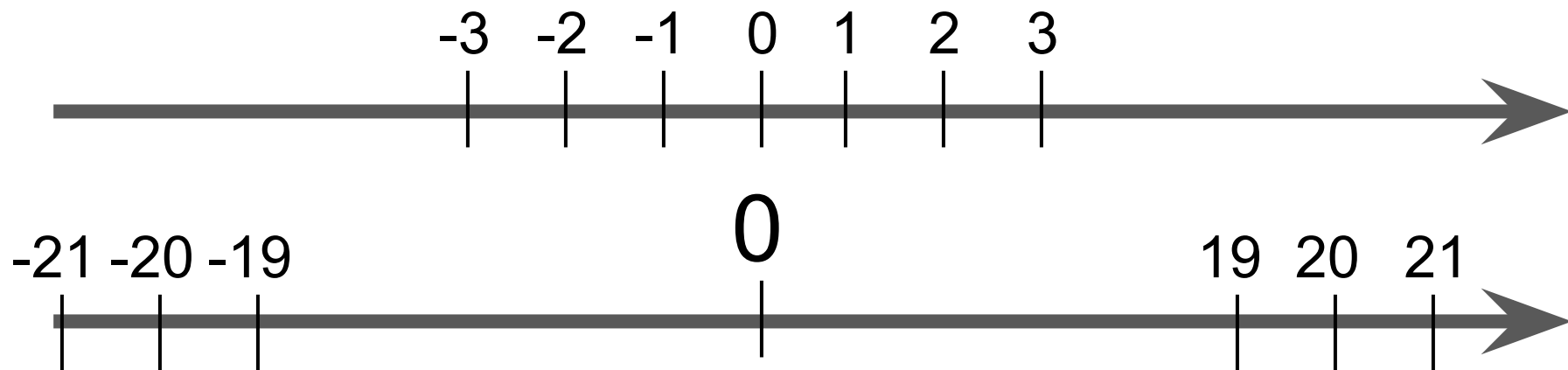
- Женщина
- 38 лет
- Android
- Слушает О. Газманова
- Оптимист

<https://journal.tinkoff.ru/citizen/>



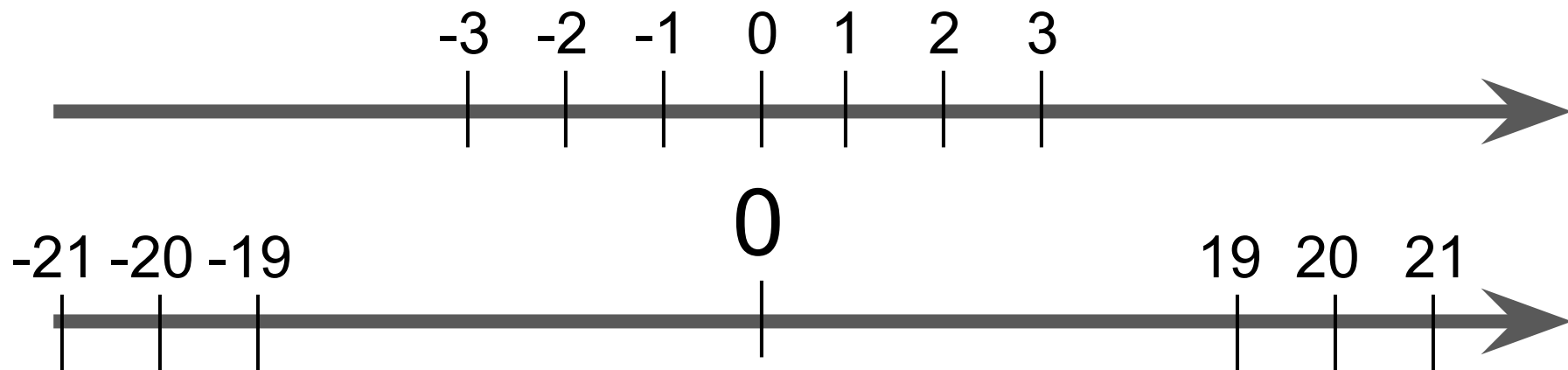
Пример №2

Вы выбираете страну для отпуска. Вот температура за 7 дней в двух странах-фаворитах:



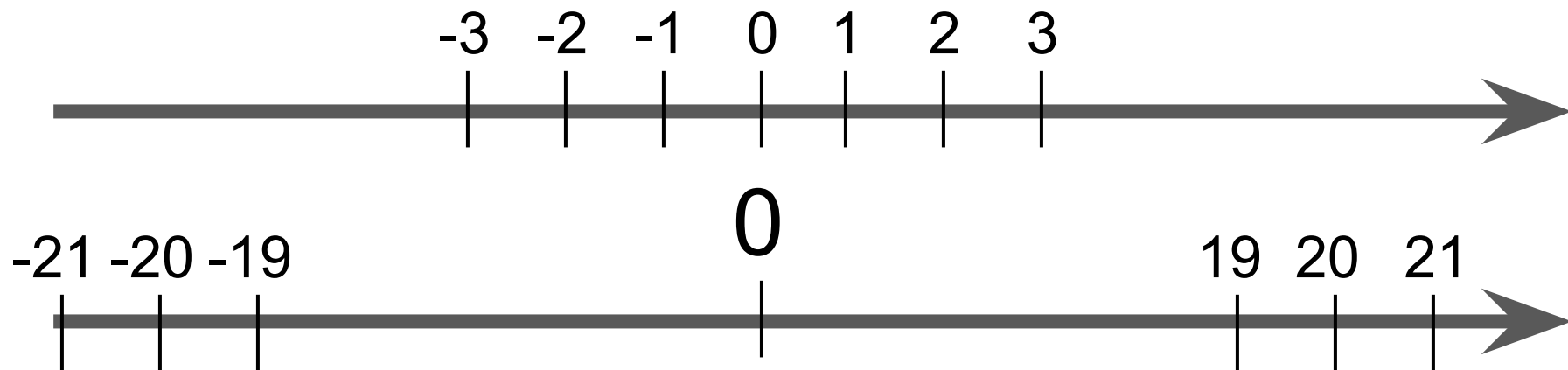
Пример №2

Каково среднее арифметическое?



Пример №2

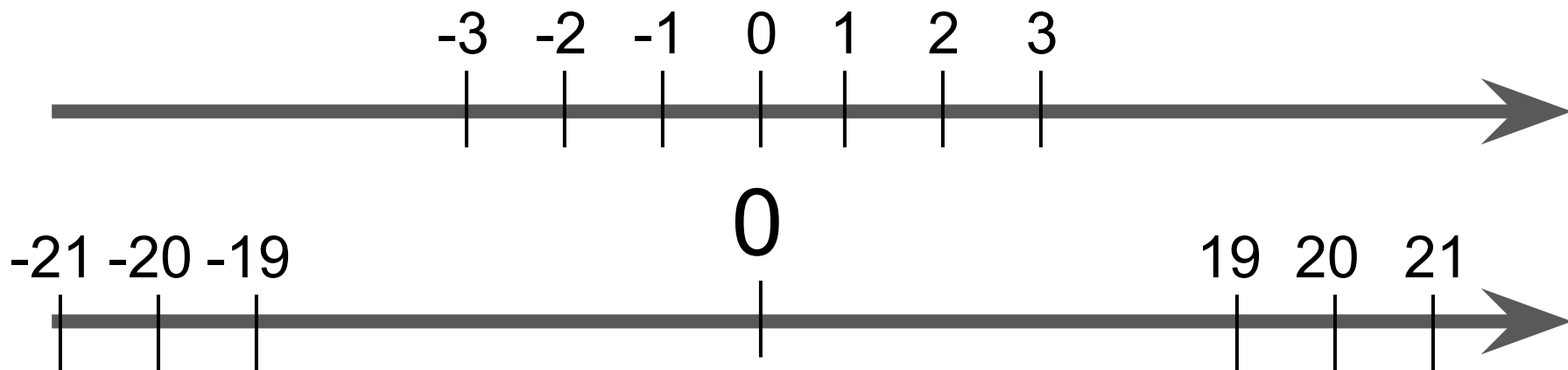
Каково среднее арифметическое? **0**



Пример №2

Каково среднее арифметическое? **0**

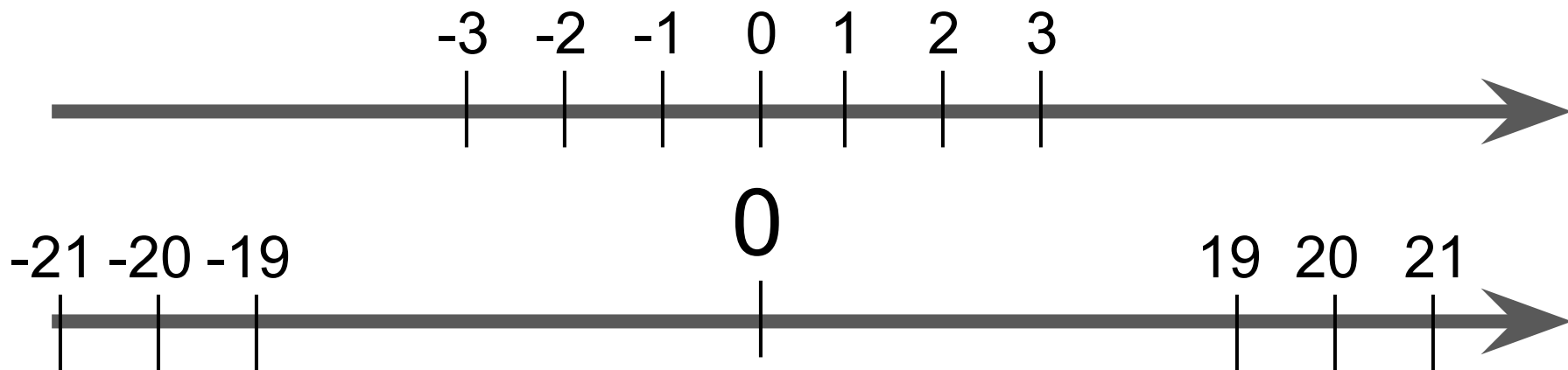
Какова медиана?



Пример №2

Каково среднее арифметическое? **0**

Какова медиана? **0**

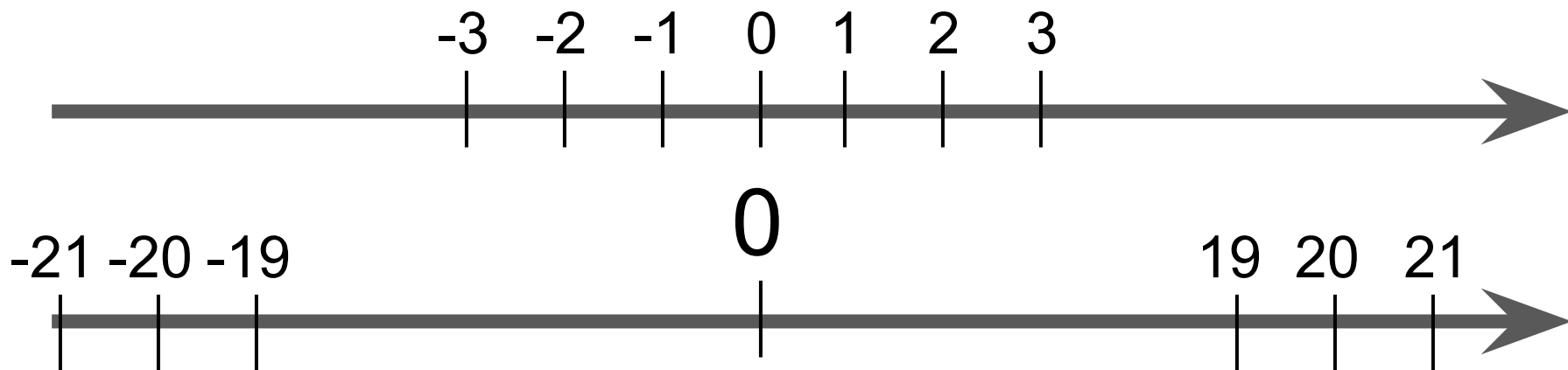


Пример №2

Каково среднее арифметическое? **0**

Какова медиана? **0**

Какую страну выберите?

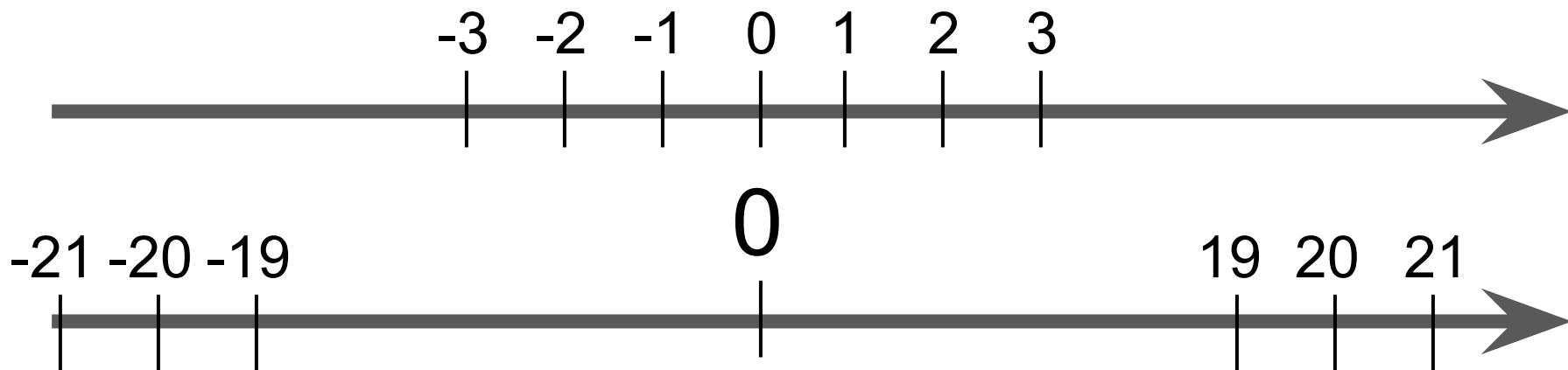


Пример №2

Каково среднее арифметическое? **0**

Какова медиана? **0**

Какую страну выберите? **Какую бы тут характеристику подобрать?**



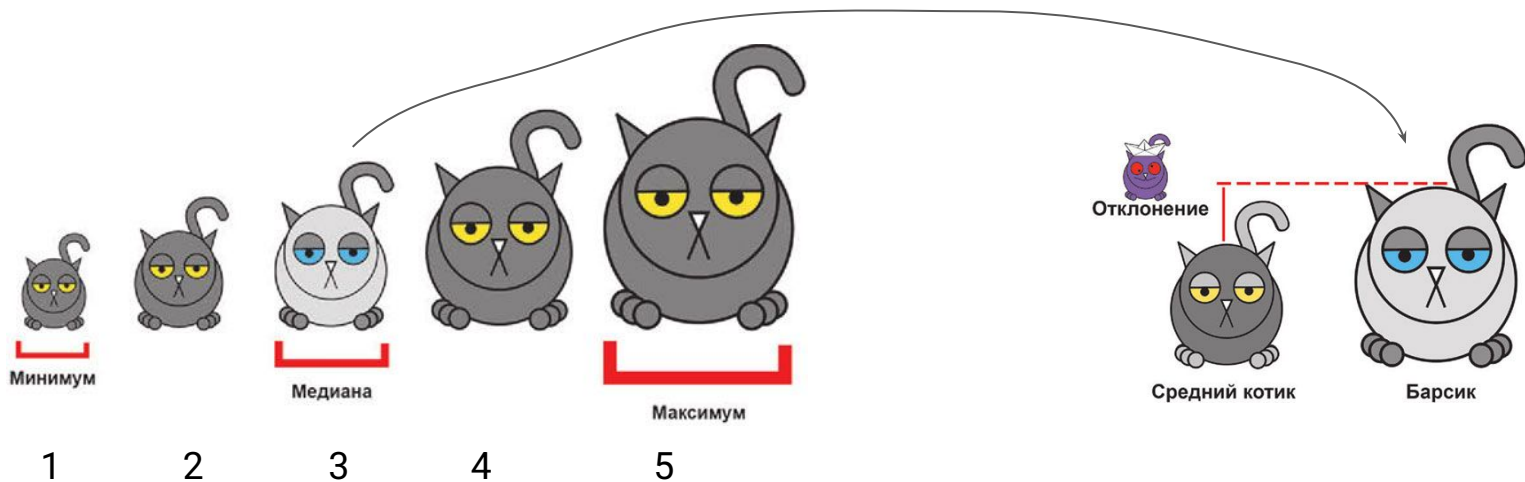
Характеристики

1. Среднее арифметическое
2. Медиана
3. **Размах**



Характеристики

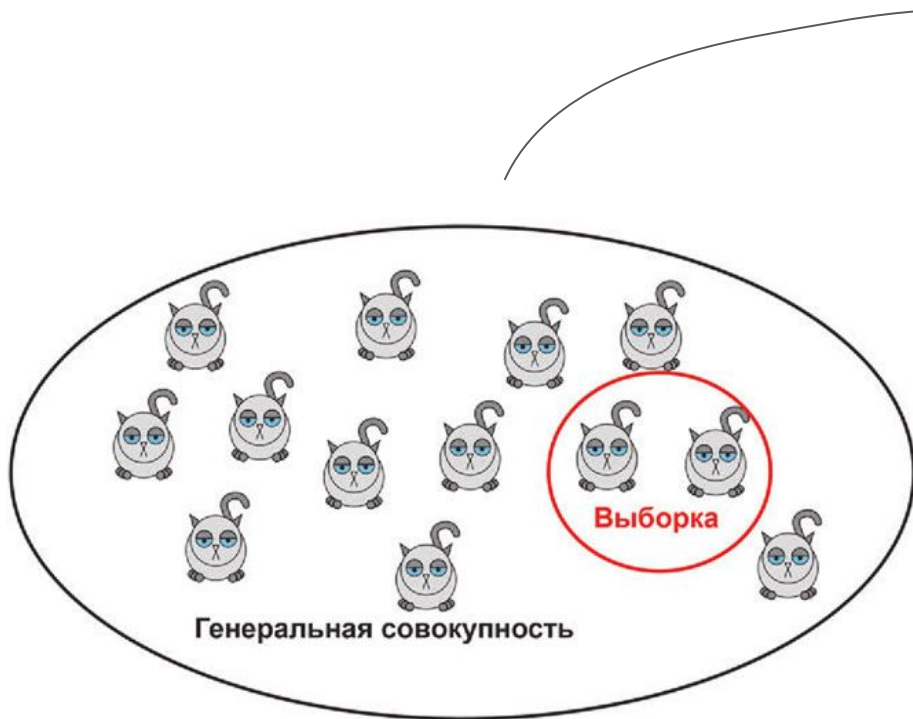
1. Среднее арифметическое
2. Медиана
3. Размах
4. **Отклонения.** Среднее отклонение считать плохо



Характеристики

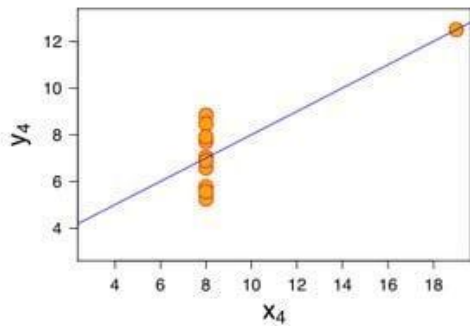
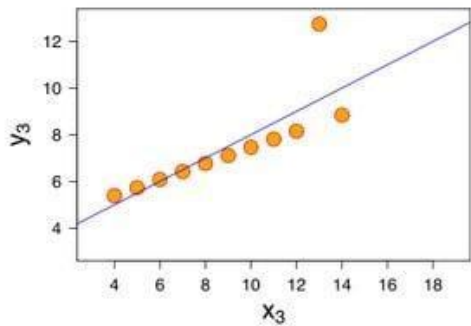
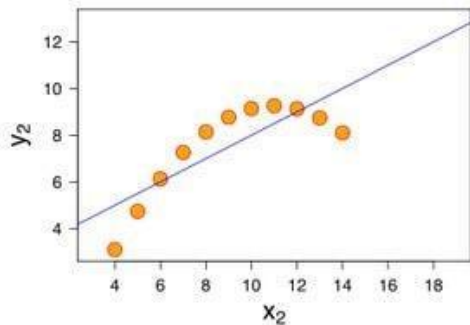
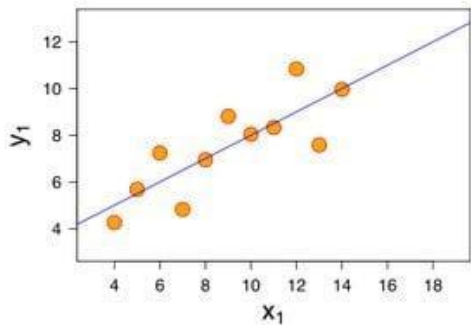
1. Среднее арифметическое
2. Медиана
3. Размах
4. **Отклонения:**
 - а. Дисперсия генеральной совокупности
 - б. Среднеквадратическое отклонение
 - с. Среднее модулей





Среднеквадратическое
отклонение

Стандартное
отклонение



Зачем нужна визуализация?

У представленных на картинке четырех наборов — одинаковые средние арифметические, медианы, дисперсии и корреляции

Plotly



- Интерактивность
- Хорошая документация
 - <https://plotly.com/>
- Просто красиво



- Front-End на **JS**
- Back-End на **Python**
- Back-End на **R**



Seaborn



Matplotlib