

Основы визуализации данных

Алексей Смагин | ВШЭ 2021



fx

год

	A	B	C	D	E
1	год	месяц	браков		
2	2006	январь	55 509		
3	2006	февраль	62 449		
4	2006	март	70 798		
5	2006	апрель	86 055		
6	2006	май	35 960		
7	2006	июнь	111 409		
8	2006	июль	127 475		
9	2006	август	149 120		
10	2006	сентябрь	151 116		
11	2006	октябрь	95 192		
12	2006	ноябрь	86 480		
13	2006	декабрь	82 101		
14	2007	январь	59 495		
15	2007	февраль	68 255		
16	2007	март	68 173		
17	2007	апрель	106 800		
18	2007	май	39 331		
19	2007	июнь	119 012		
20	2007	июль	147 253		
21	2007	август	163 630		

Зачем нужна визуализация?

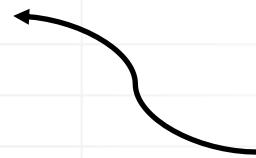


fx | год

	A	B	C	D	E
1	год	месяц	браков		
2	2006	январь	55 509		
3	2006	февраль	62 449		
4	2006	март	70 798		
5	2006	апрель	86 055		
6	2006	май	35 960		
7	2006	июнь	111 409		
8	2006	июль	127 475		
9	2006	август	149 120		
10	2006	сентябрь	151 116		
11	2006	октябрь	95 192		
12	2006	ноябрь	86 480		
13	2006	декабрь	82 101		
14	2007	январь	59 495		
15	2007	февраль	68 255		
16	2007	март	68 173		
17	2007	апрель	106 800		
18	2007	май	39 331		
19	2007	июнь	119 012		
20	2007	июль	147 253		
21	2007	август	163 630		

Зачем нужна визуализация?

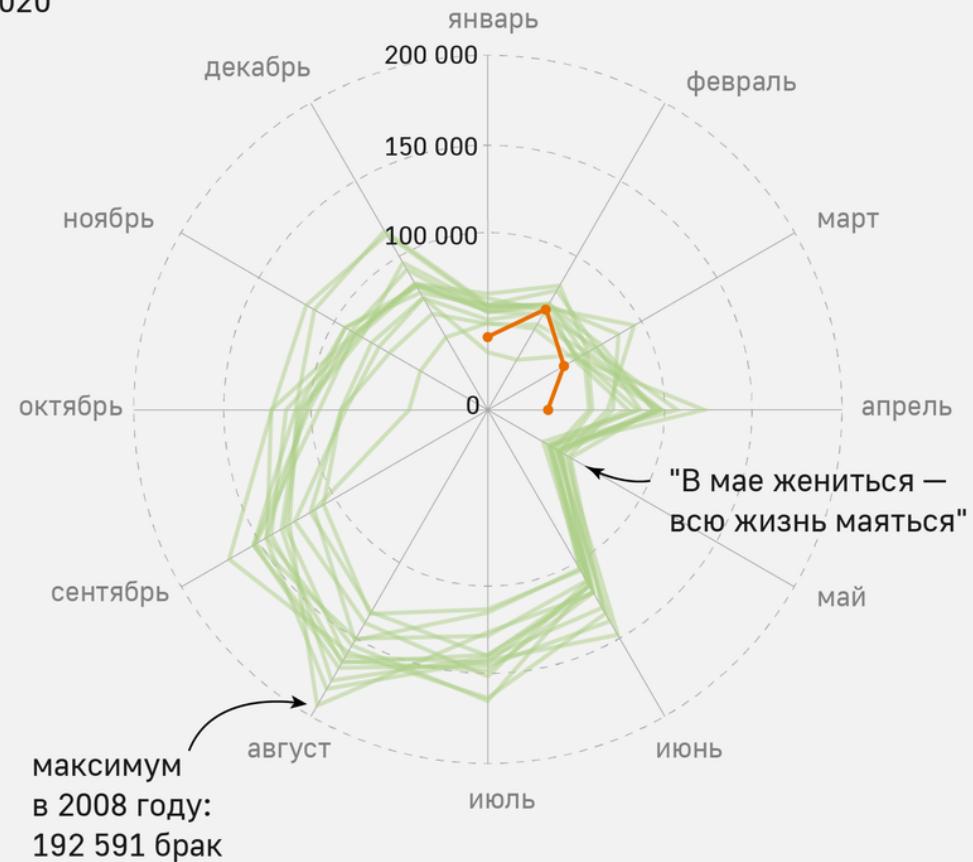
Таблица идеально подходит,
если нужно рассмотреть
каждое значение



Число браков в России

2006-2019

2020

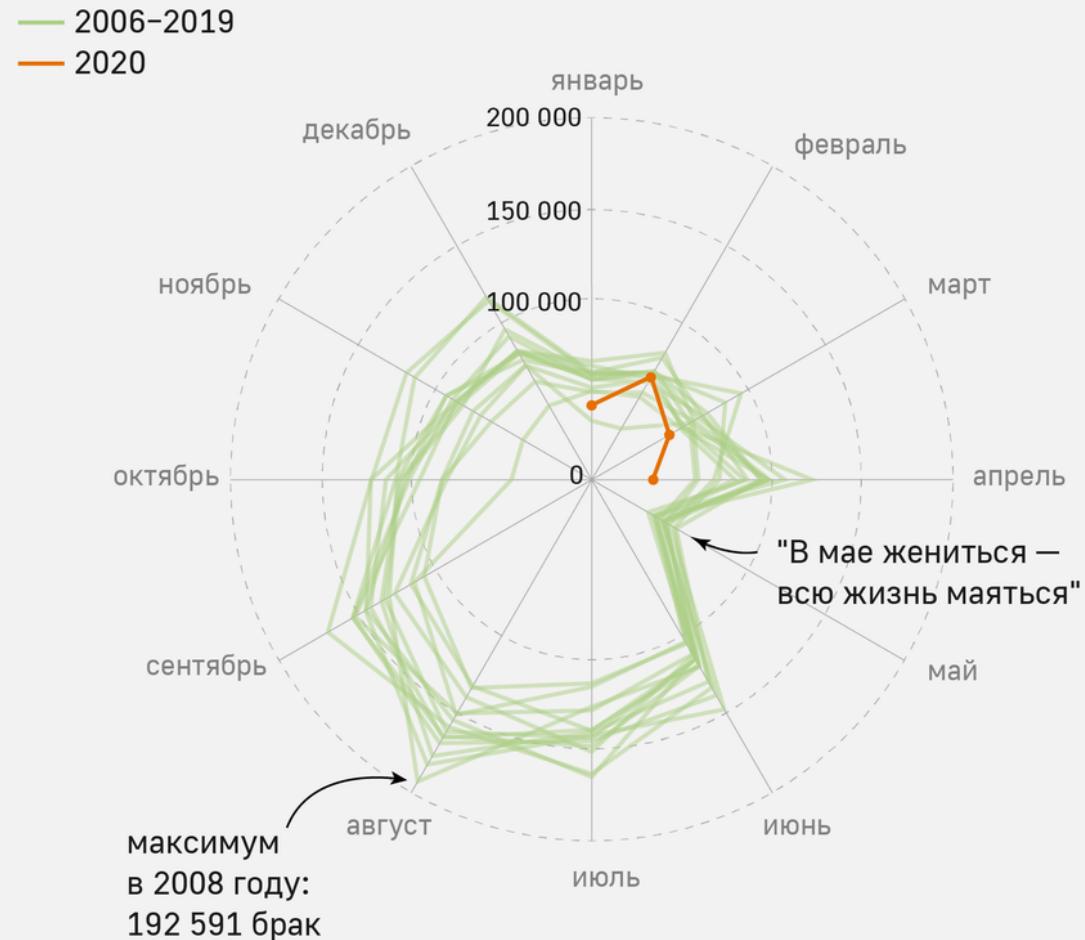


**Визуализация
помогает наглядно донести
ключевые идеи, которые
находятся в данных**

Элементы диаграммы

Число браков в России ← заголовок

— заголовок



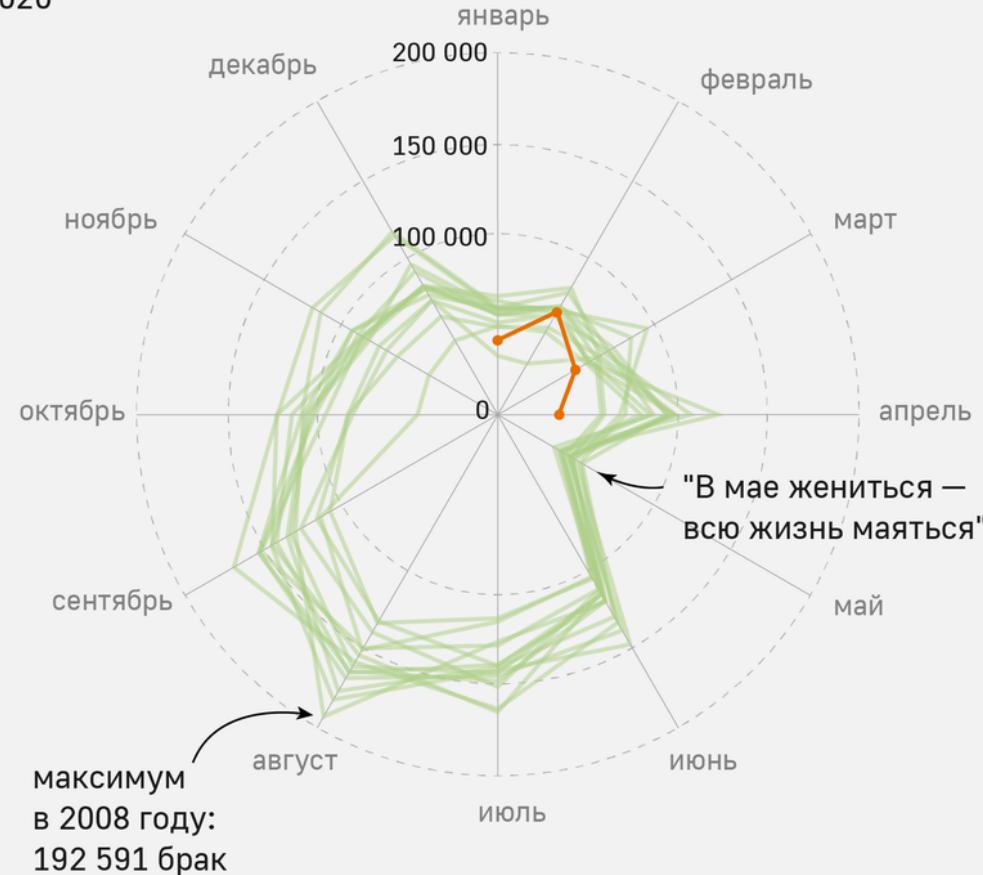
© ТАСС, 2020. Источник: Росстат.

Число браков в России

заголовок

легенда

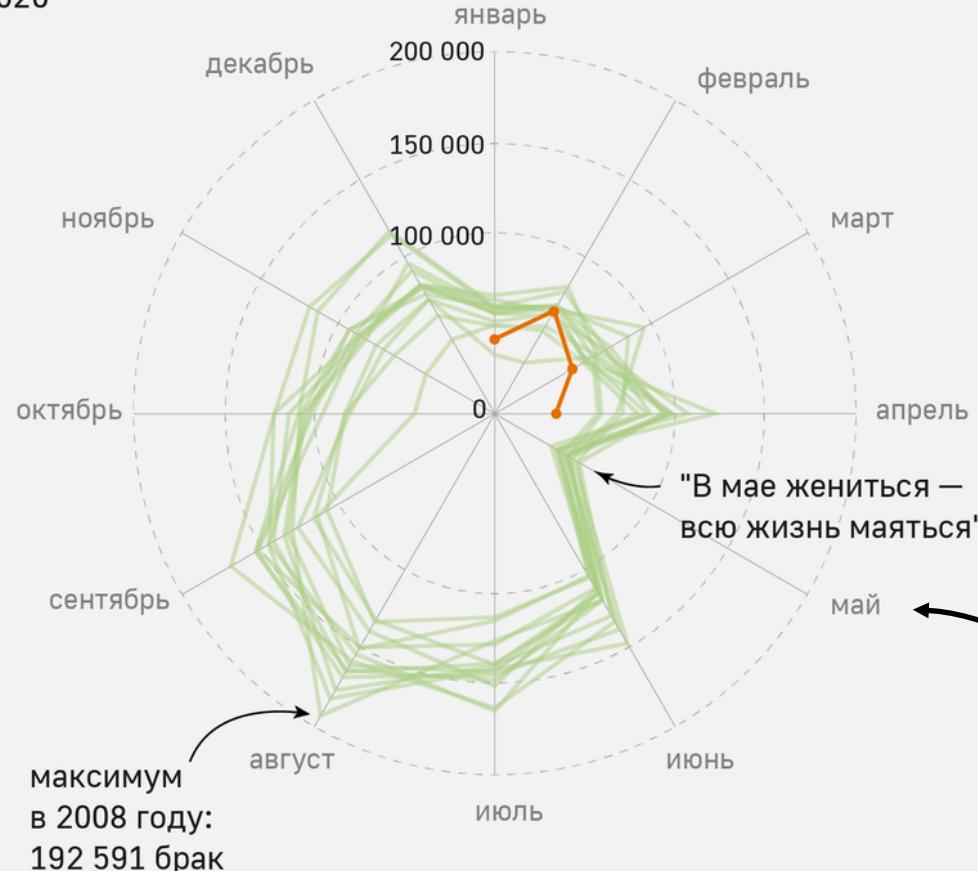
- 2006-2019
- 2020



Число браков в России

легенда →

2006-2019
2020



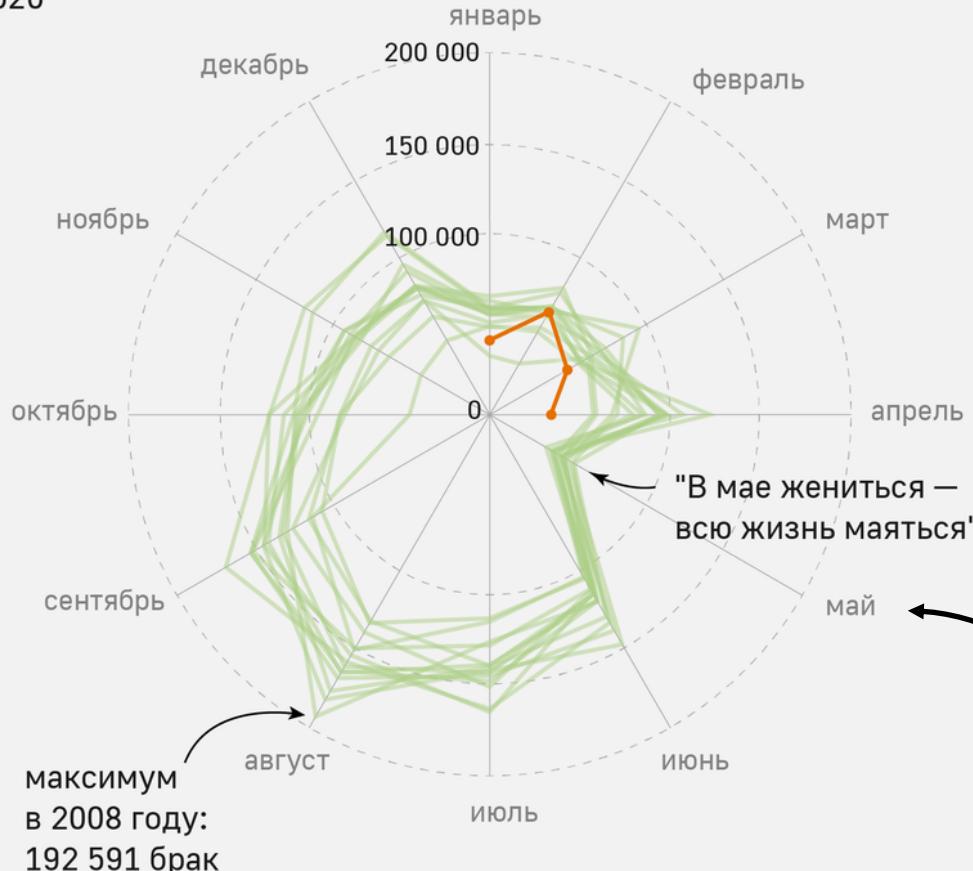
оси и
числовые
подписи

Число браков в России

легенда

2006-2019
2020

аннотации



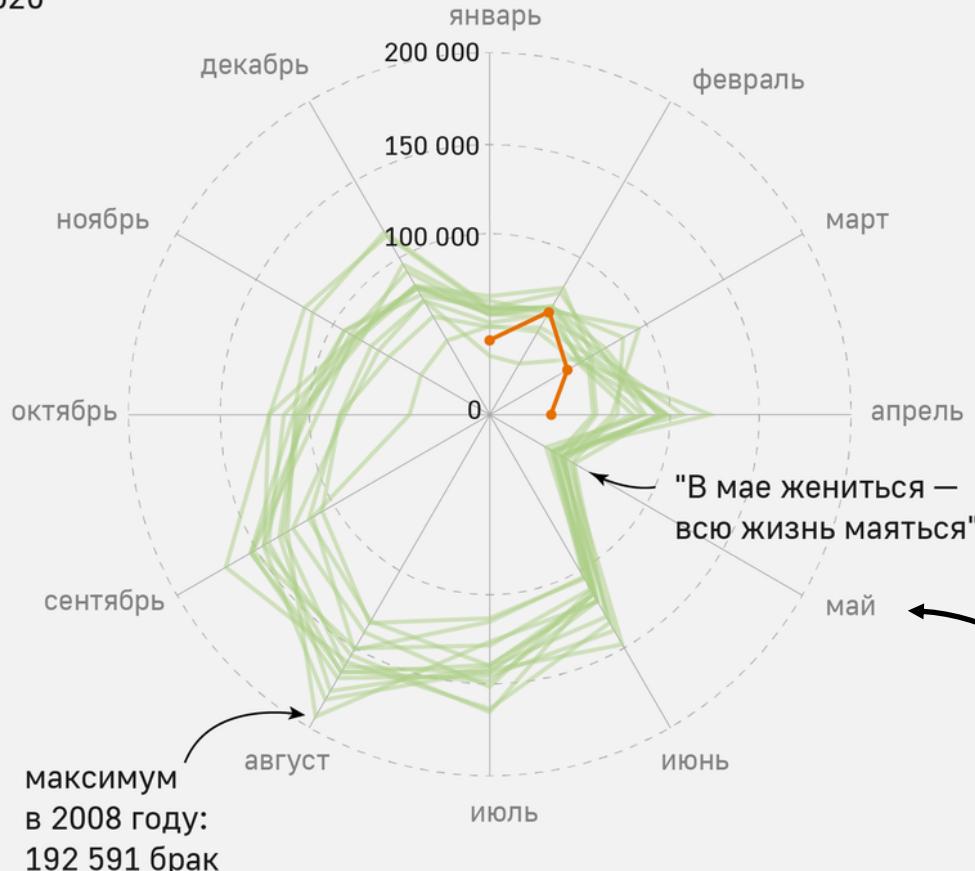
оси и
числовые
подписи

Число браков в России

легенда

2006-2019
2020

аннотации



© ТАСС, 2020. Источник: Росстат.

источник и авторы

оси и
числовые
подписи

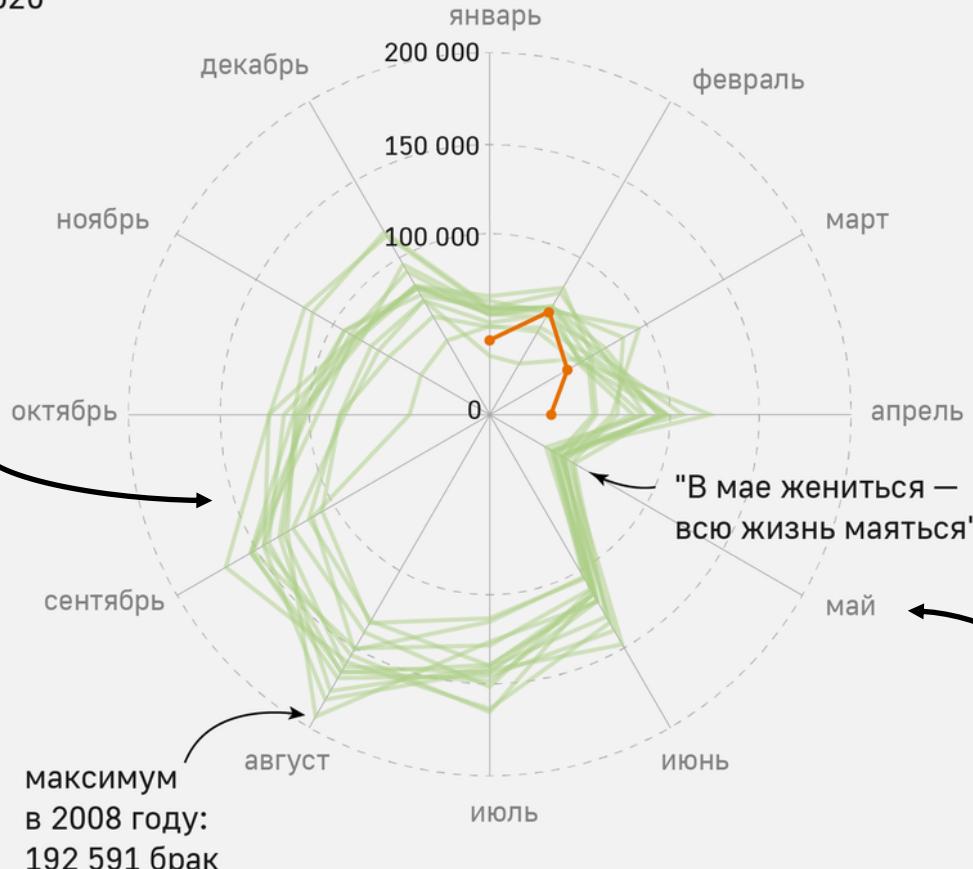
Число браков в России

легенда

2006-2019
2020

закодированная
информация

аннотации



оси и
числовые
подписи

© ТАСС, 2020. Источник: Росстат.

источник и авторы

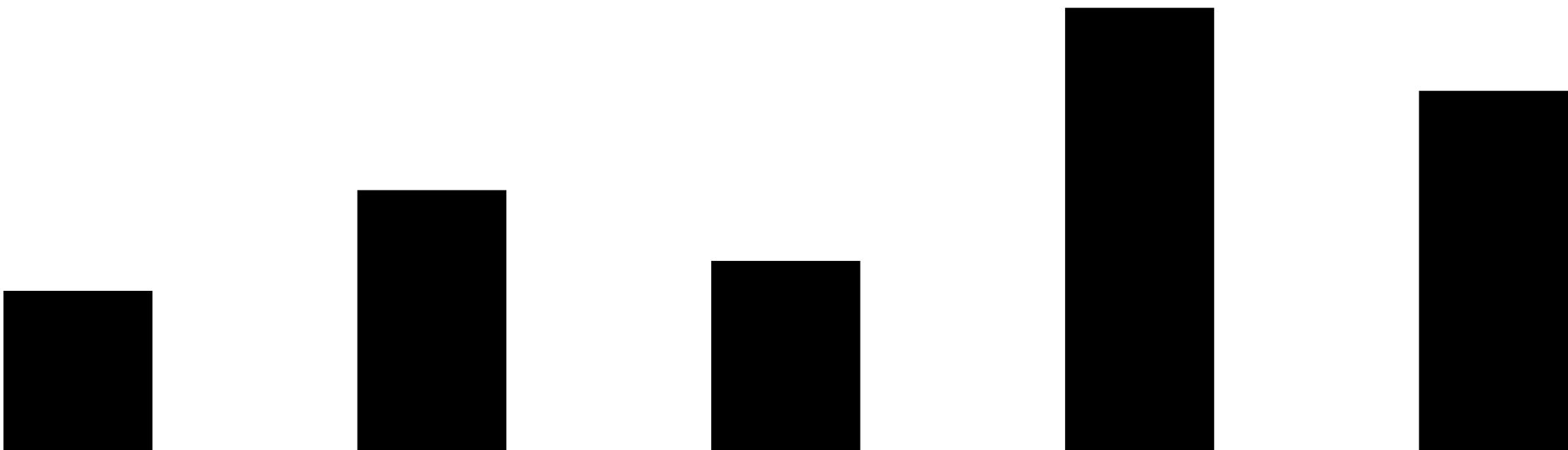
Кодирование

Всем рулят фигуры и их свойства

Каждая фигурка —
это какая-то сущность



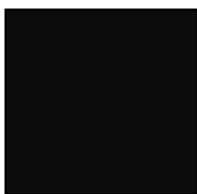
можно менять их размер



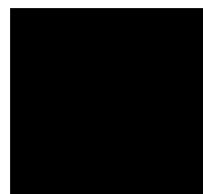
ЦВЕТ



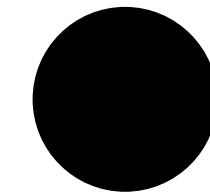
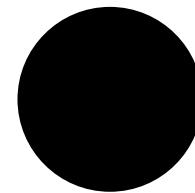
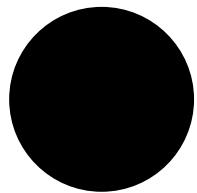
насыщенность



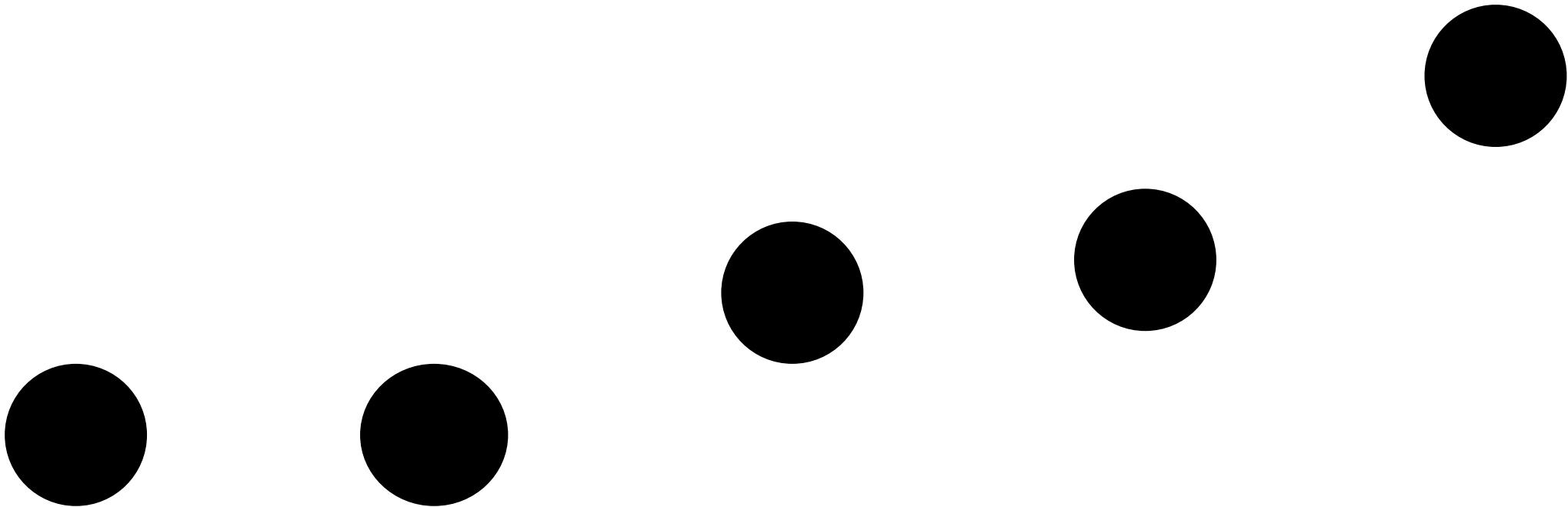
группировать их



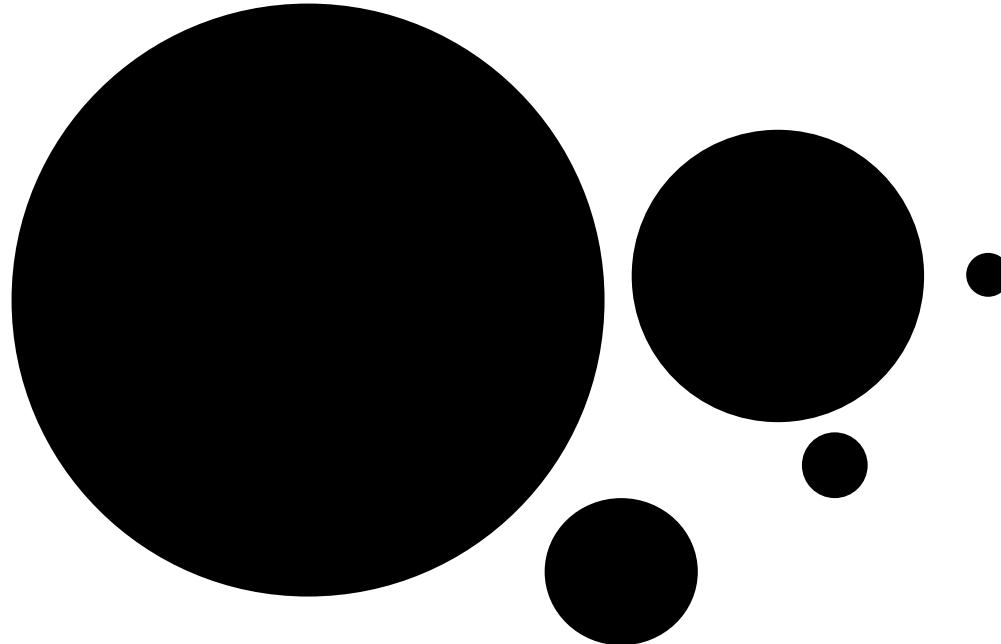
придавать любую форму



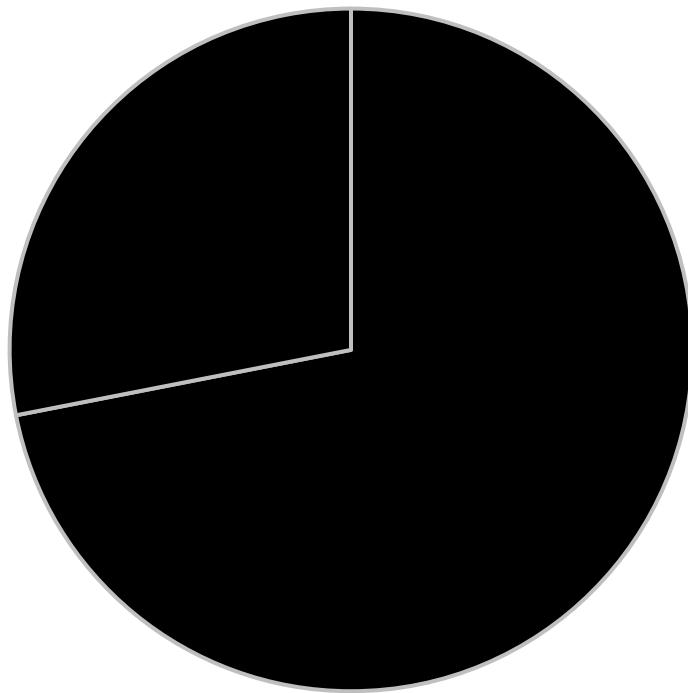
менять положение



или сразу несколько свойств



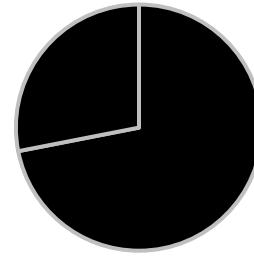
делить на части



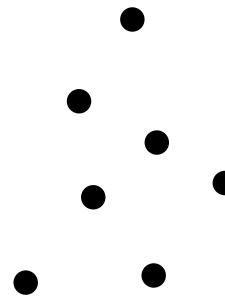
Свойства объектов кодируем свойствами фигур



высота



угол



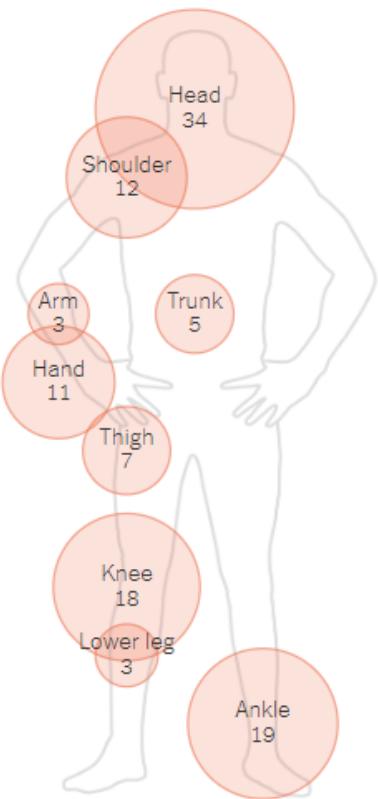
координаты

**Как
закодирована
информация?**

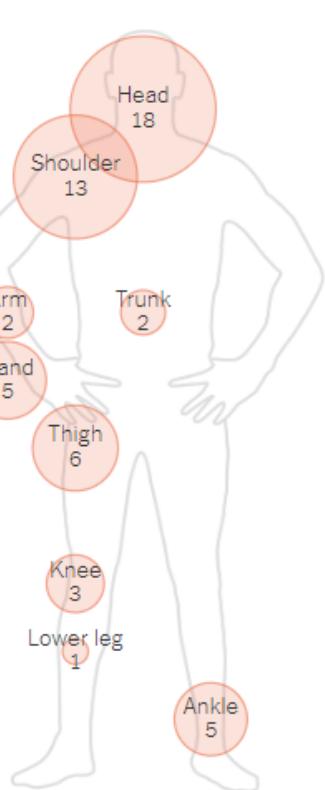
Common injuries for boys among popular high school sports

Injuries per 10,000 competition plays

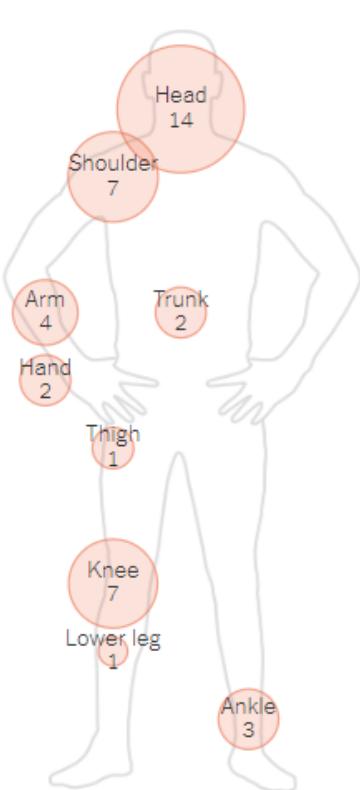
Football



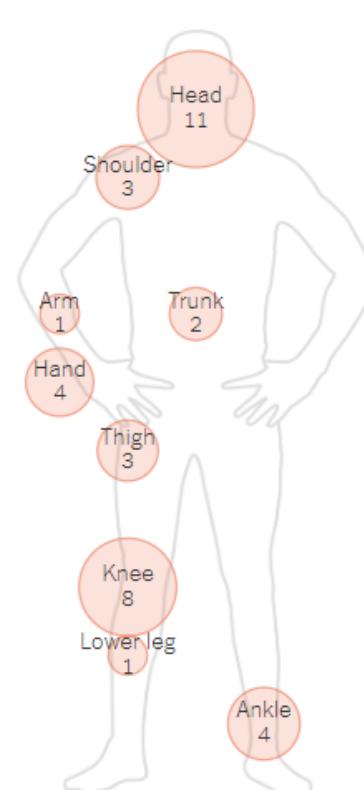
Hockey



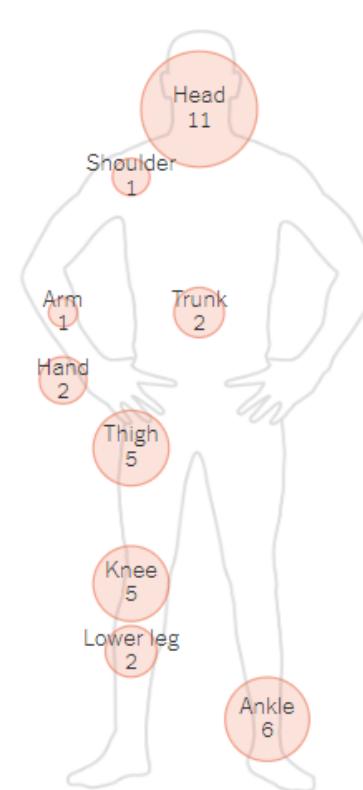
Wrestling



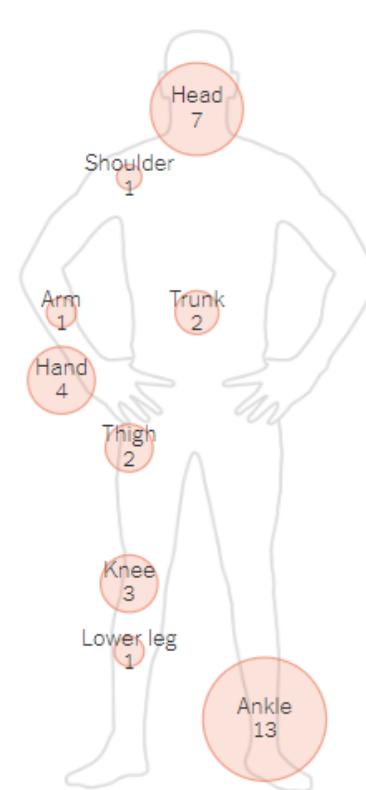
Lacrosse



Soccer



Basketball



Где живут российские семьи 

■ Собственное жильё ■ Аренда ■ Общежитие



2000

2017

Карта заражения коронавирусом

<

29 октября

>

▶

II.

Заболели

1 586 255
+ 17 743 (1.1%)
за деньСПб
59 721Мурмн
19 457Чукот
239Камчат
5 672Москва
413 928Карел
6 402

Выздоровели

1 190 665
+ 13 535 (1.1%)
за деньКлинггрд
6 634Лен обл
11 179Новг
7 187Влгда
6 818НАО
494ЯНАО
20 190Крск
25 860Магад
4 357Псков
7 727Тверь
9 276Ярсвл
11 164Иванво
11 246Кстрма
7 104Мар Эл
6 029Архнгл
19 929Коми
12 926ХМАО
27 877Тюмень
13 105Томск
11 977Кузбас
15 057Ирк
24 378Смлнск
8 578Калуга
11 731Мск обл
85 762Влдмр
9 407Нижний
40 768Чуваш
9 782Татар
8 328Удмрт
8 868Екб
35 650Курган
5 723Нск
16 939Хакас
6 675Бурят
11 258Брянск
12 481Орел
11 721Тула
12 131Рязань
10 519Мрдв
8 121Ульнск
20 592Самара
15 625Бшкр
10 586Алтай
7 271Амур
6 331Прмрск
14 980

Летальность

3.4%

Летальность считается как
отношение числа умерших
к сумме умерших и выздоровевших

Курск 10 652	Липецк 7 937	Тамбов 9 660	Пенза 14 920	Сртв 20 122	Орнбрг 17 210
-----------------	-----------------	-----------------	-----------------	----------------	------------------

Белг 11 350	Врнж 24 477	Влггрд 19 919
----------------	----------------	------------------

Свстпл 2 238	Крым 9 172	Адыгея 5 701	Кубань 17 119	Ростов 29 995	Клмк 8 904	Астрхн 9 485
-----------------	---------------	-----------------	------------------	------------------	---------------	-----------------

КЧР 9 437	Ставр 20 890	Чечня 2 617	Дгстн 15 875
--------------	-----------------	----------------	-----------------

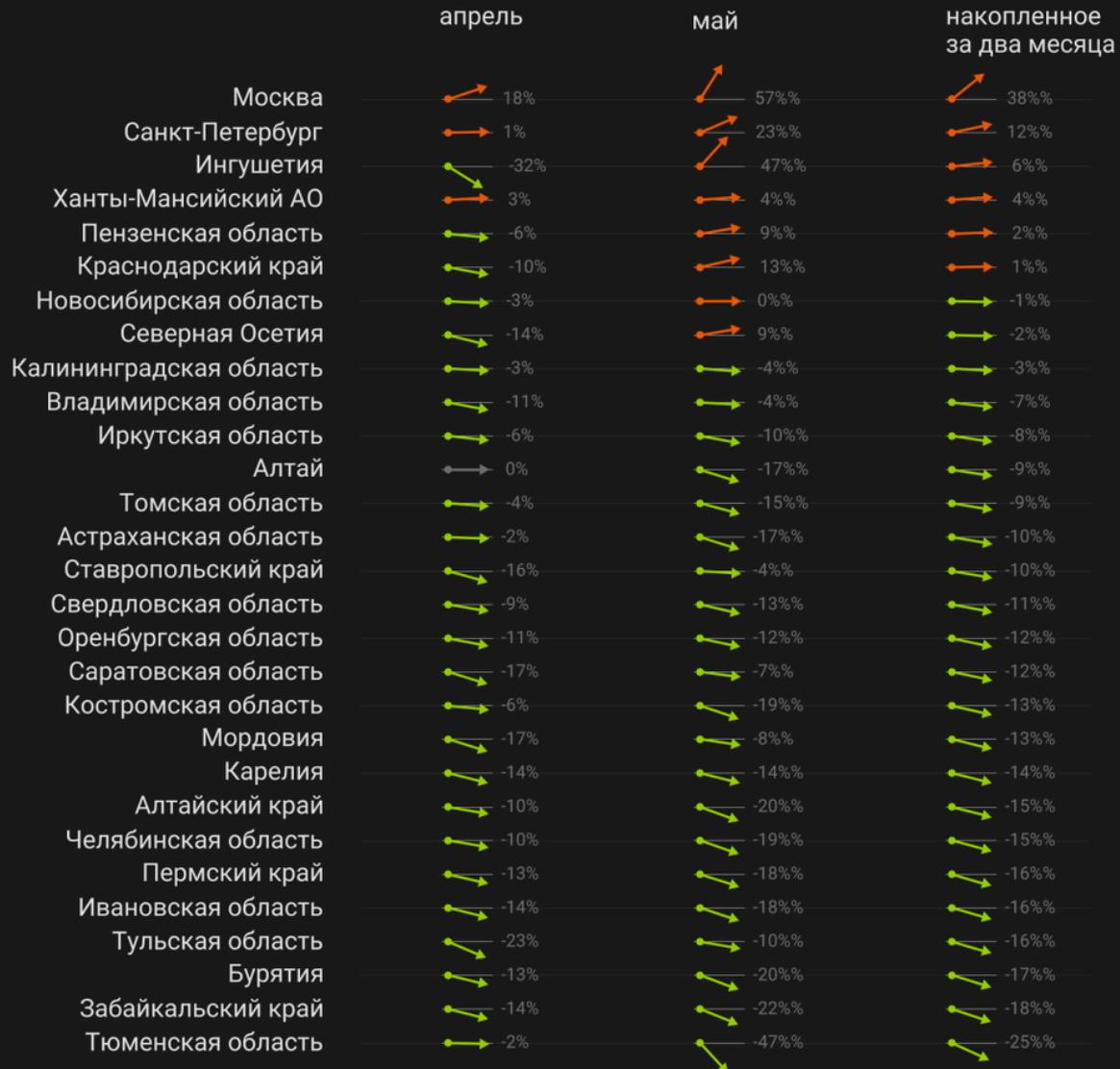
КБР 9 921	Алания 7 154	Ингуш 7 225
--------------	-----------------	----------------

Медиазона

Официальные данные федеральных
и региональных российских властей
на 29 октября 2020 года

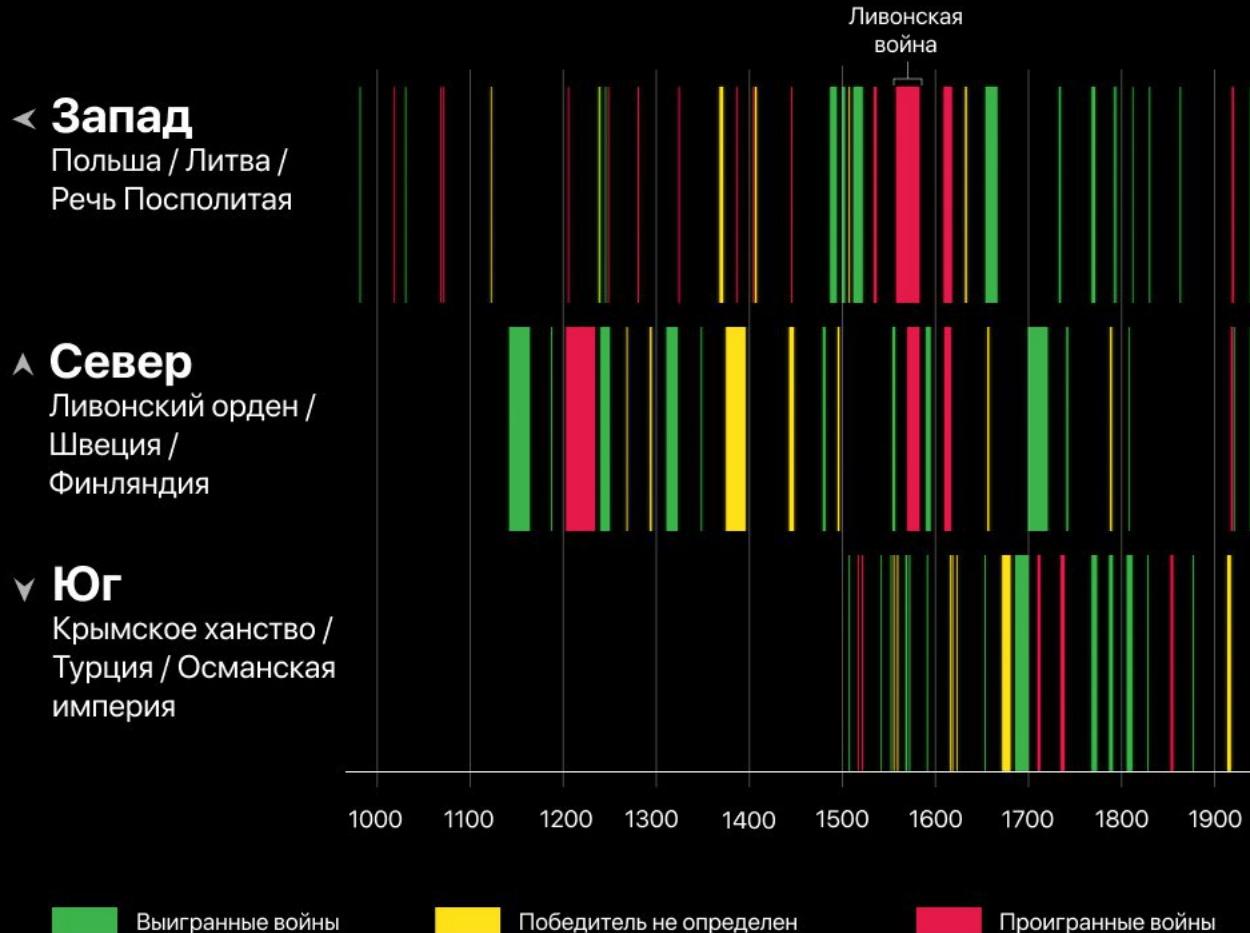
Избыточная смертность по итогам апреля и мая* зафиксирована в 6 регионах

*смертность за май известна только для 29 субъектов Федерации



Войны России

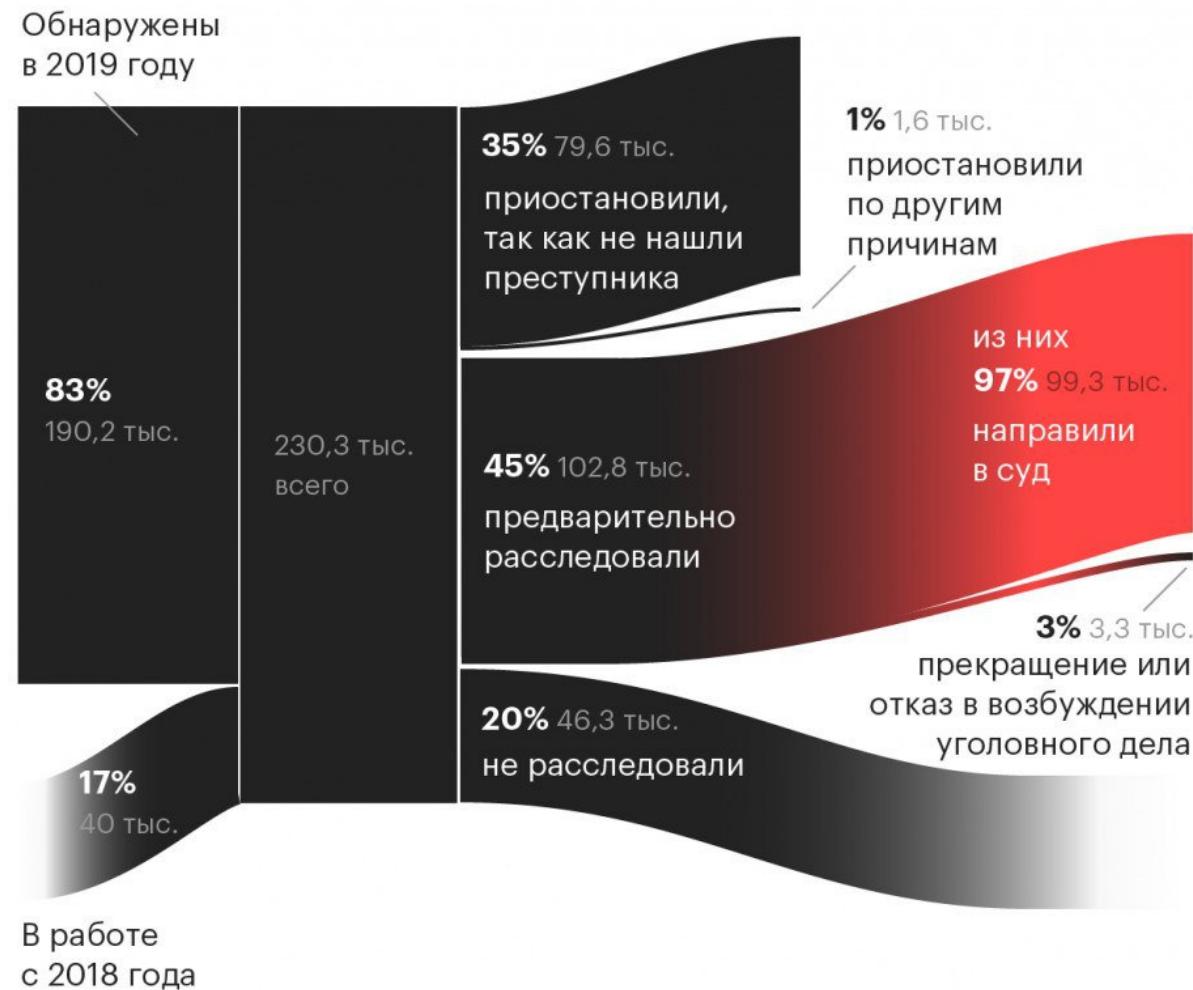
с главными европейскими конкурентами, 982–1941



Правопреемники, датировка, участники и победители войн — по историографии,
сложившейся в русской Википедии летом 2019 года.

Что происходит с делами об обороте наркотиков в России

Число преступлений по «наркотическим» статьям Уголовного кодекса РФ, бывших в работе в 2019 году



Отвратительные графики

прямой эфир

Динамика выявленных случаев заражения COVID-19 в России



Общее число заражённых за весь период эпидемии – **353 427**

Источник: Оперативный штаб

РБК 14:01 мск 3 тыс. специалистов rbc.ru Мишустин: выплаты медраб
EUR ЦБ 78,44 +0,83% дитный рынок RUONIA 5,32 -0,746% Мосбиржа АЛРО



Why you should begin investing at birth

hypothetical 6% annual return

Begin at birth, \$5
a day until 25:

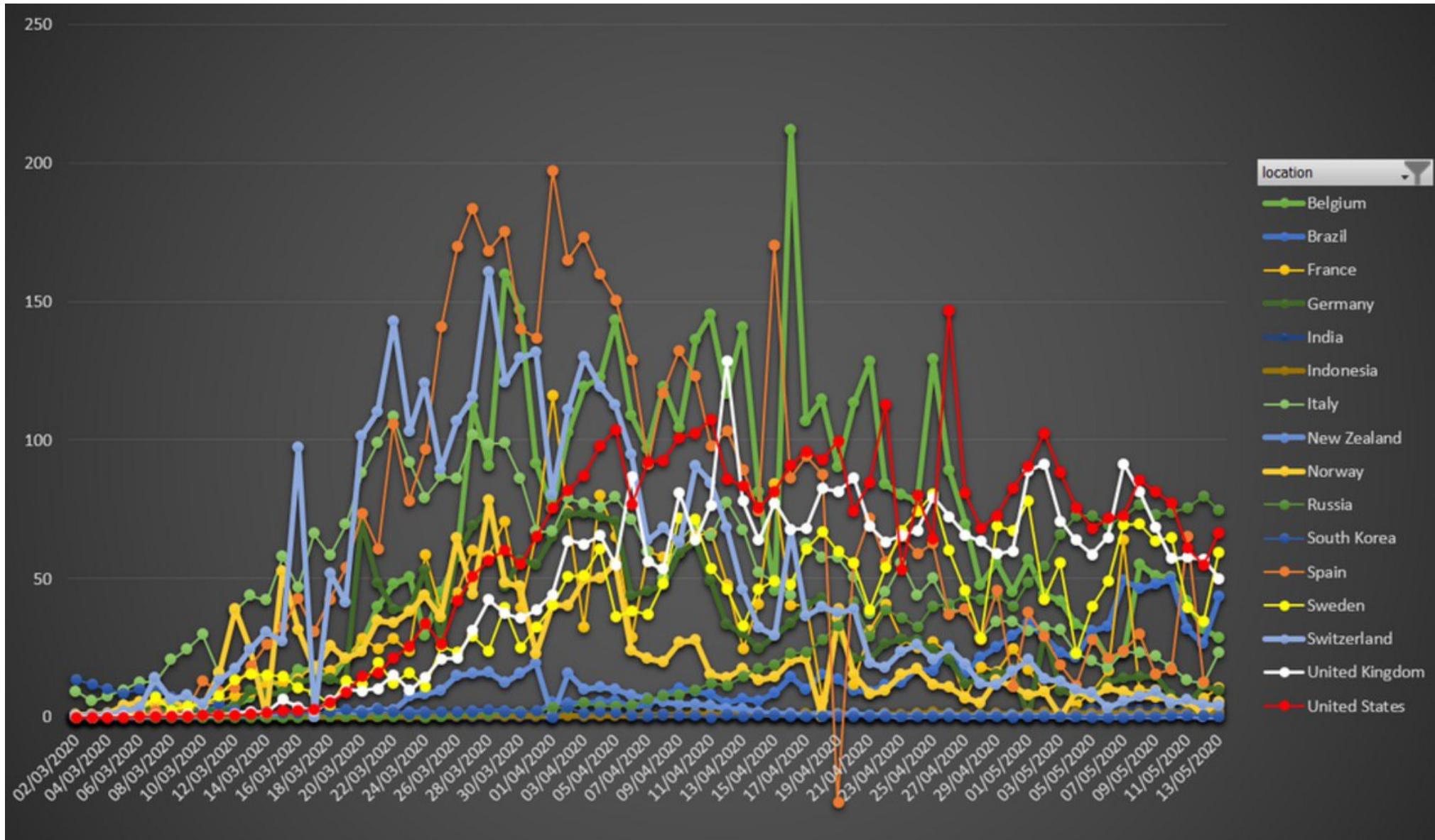
\$106,208



Begin at 18, \$5
a day until 25:

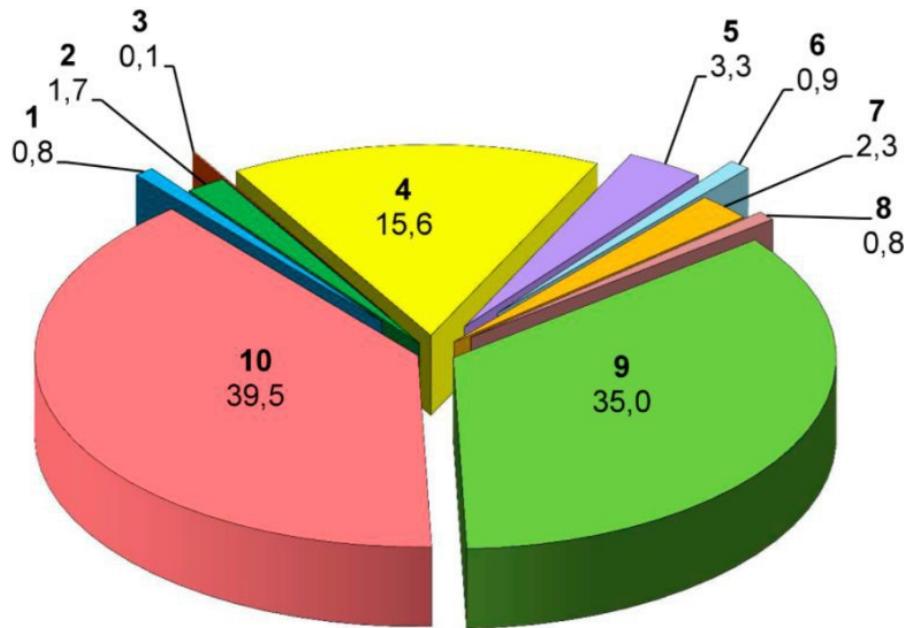
\$16,249





СТРУКТУРА ПРЕСТУПНОСТИ (в %)

январь - июнь



-
- 1 - взяточничество
2 - убийство, умышленное причинение тяжкого вреда здоровью, изнасилование
3 - хулиганство
4 - мошенничество

Признаки хорошего графика:

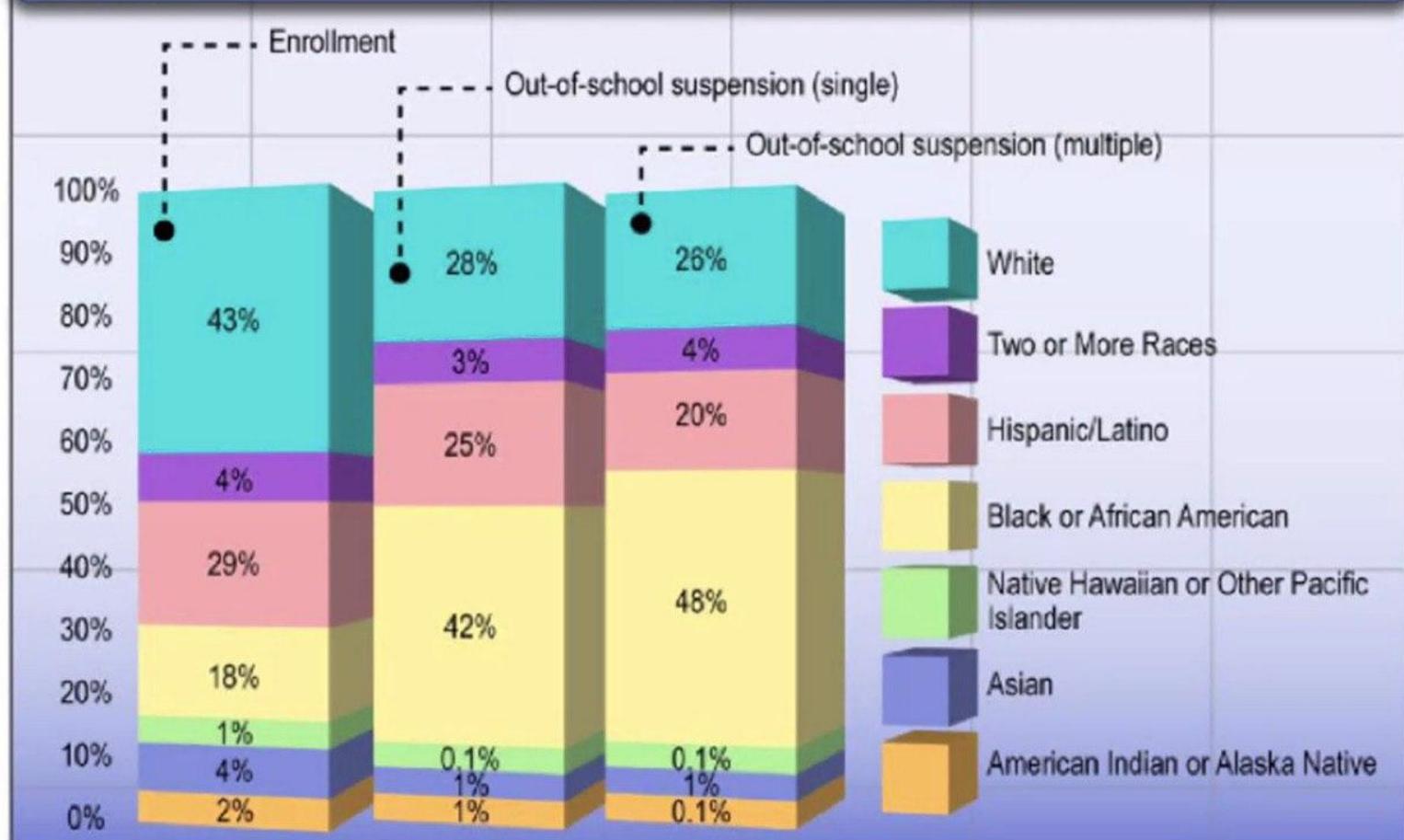
1. Он основан на данных, вызывающих доверие
2. Он корректно отображает информацию
3. Он понятен и чётко доносит идею
4. Он выглядит привлекательно

Зачем мы делаем график?

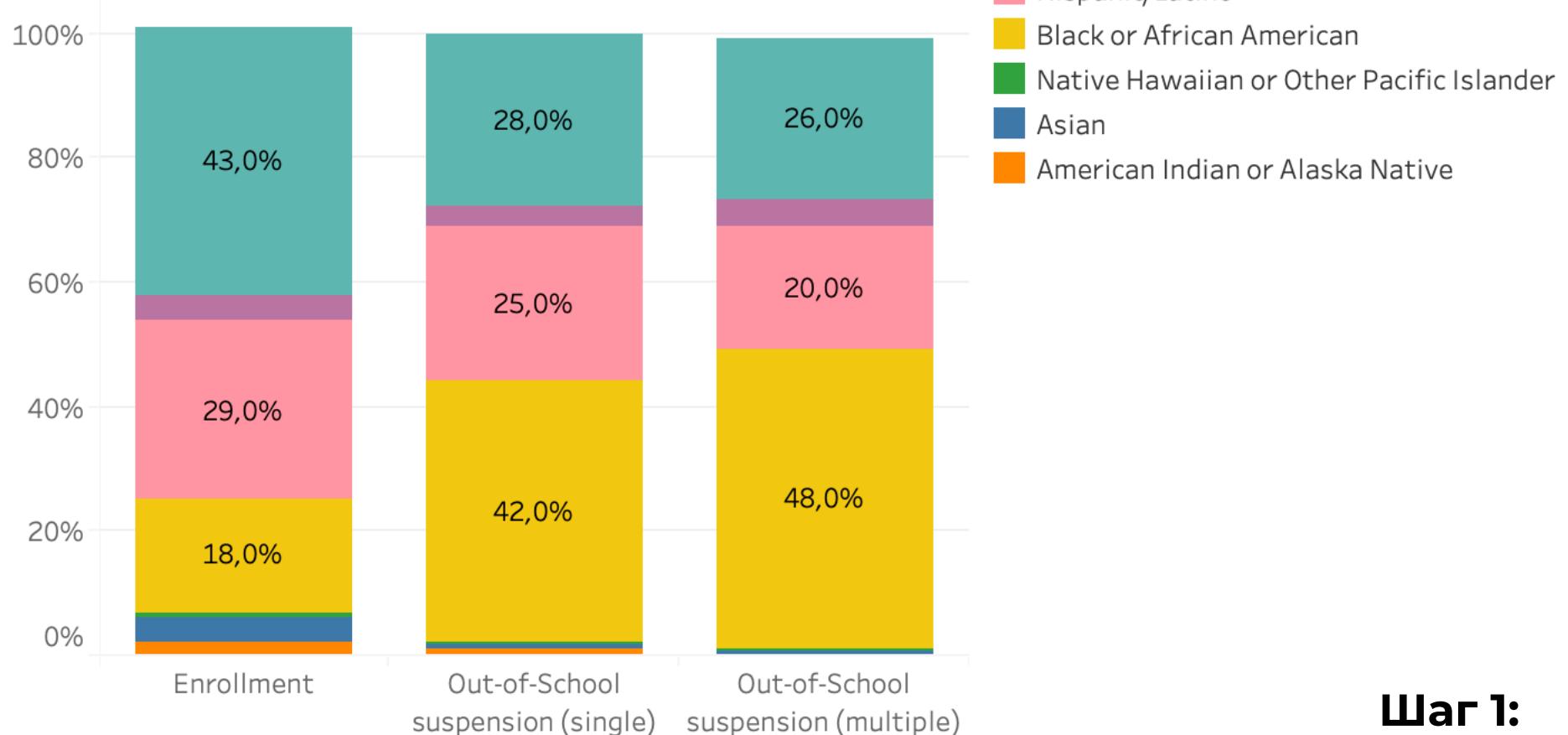
- 1. Для собственного исследования**
- 2. «Поисковая визуализация»**
- 3. Презентация**
- 4. Научная публикация**
- 5. Отчёт**
- 6. Визуализация в СМИ**

**Что делать с
плохим
графиком?**

Preschool students receiving suspensions, by race and ethnicity

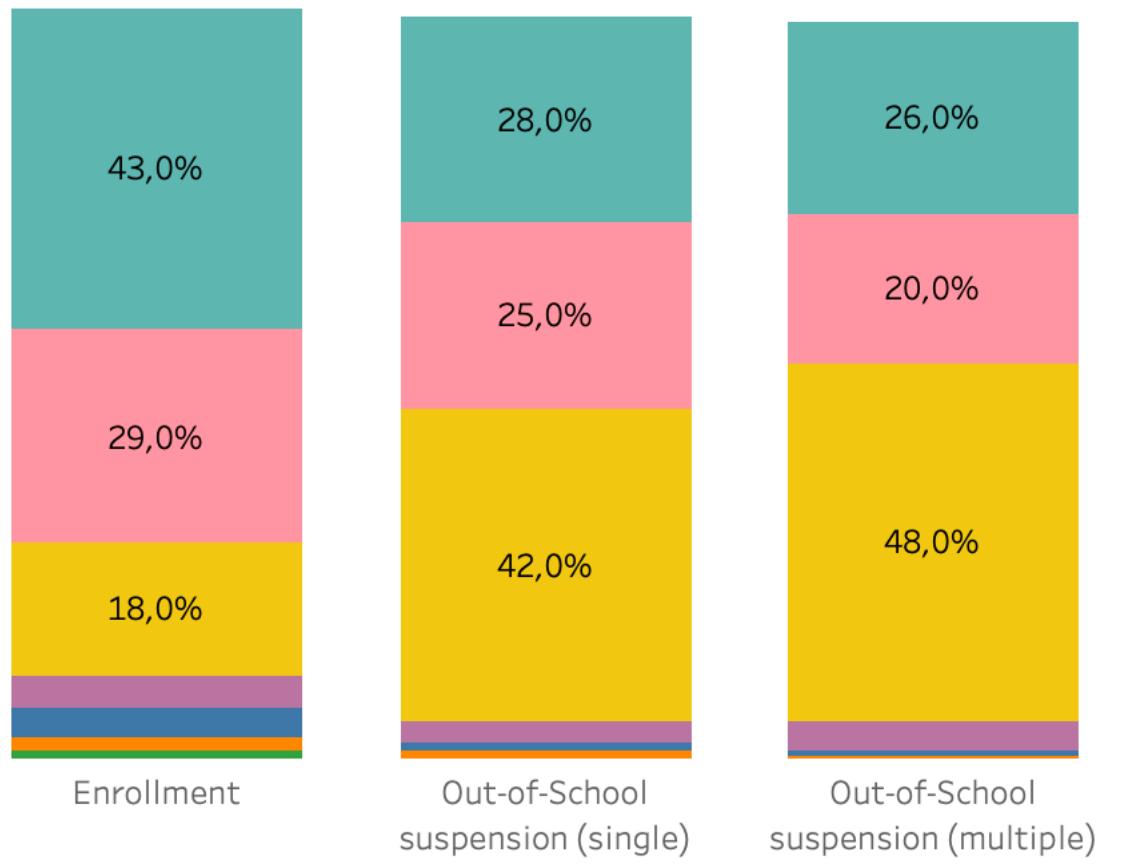


Preschool students receiving suspensions, by race and ethnicity



Шаг 1:
Убираем трёхмерность
и градиентный фон

Preschool students receiving suspensions, by race and ethnicity

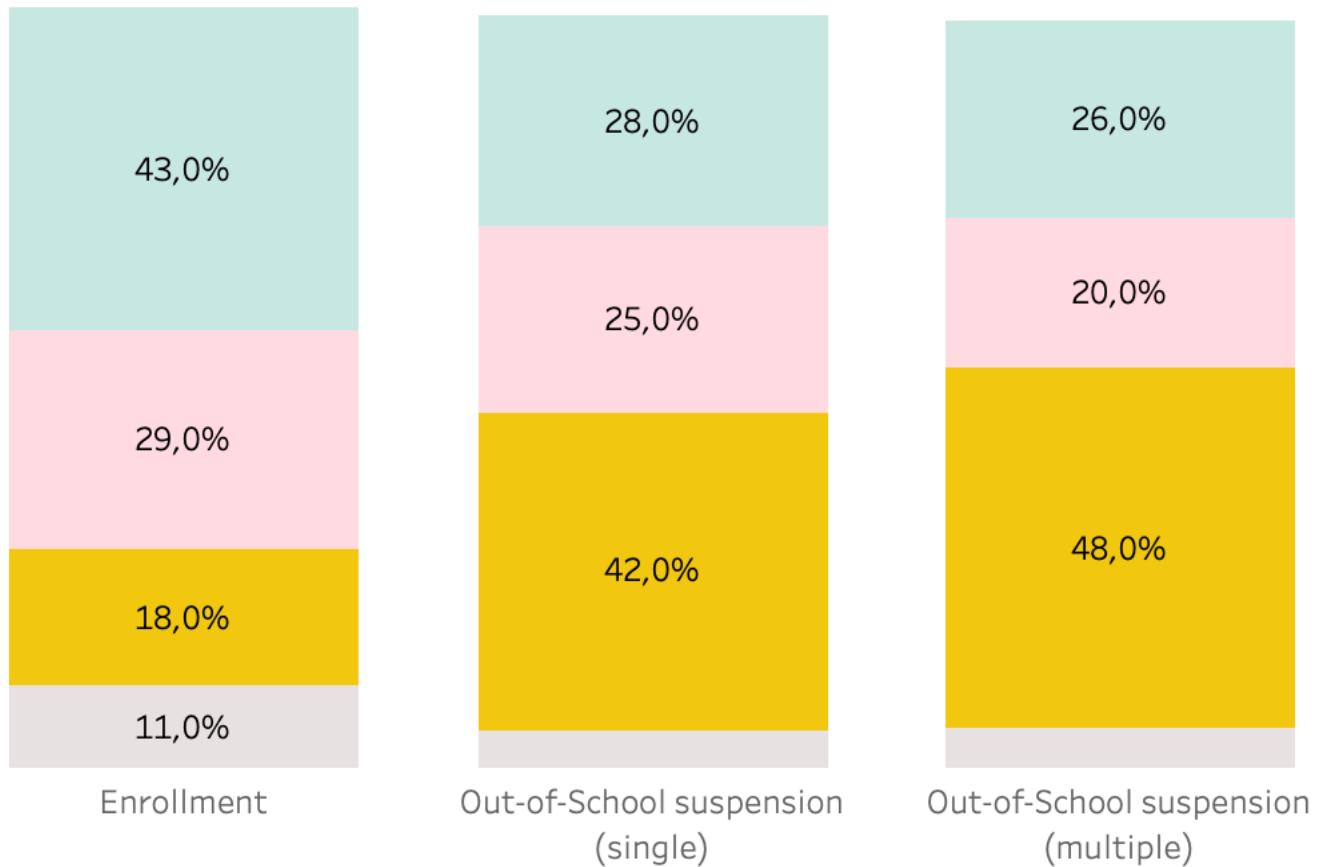


Race

- White
- Hispanic/Latino
- Black or African American
- Two or more races
- Asian
- American Indian or Alaska Native
- Native Hawaiian or Other Pacific Islander

Шаг 2:
Делаем осмысленную
сортировку, убираем
ось и сетку

Black preschool students are more likely suspended than other children

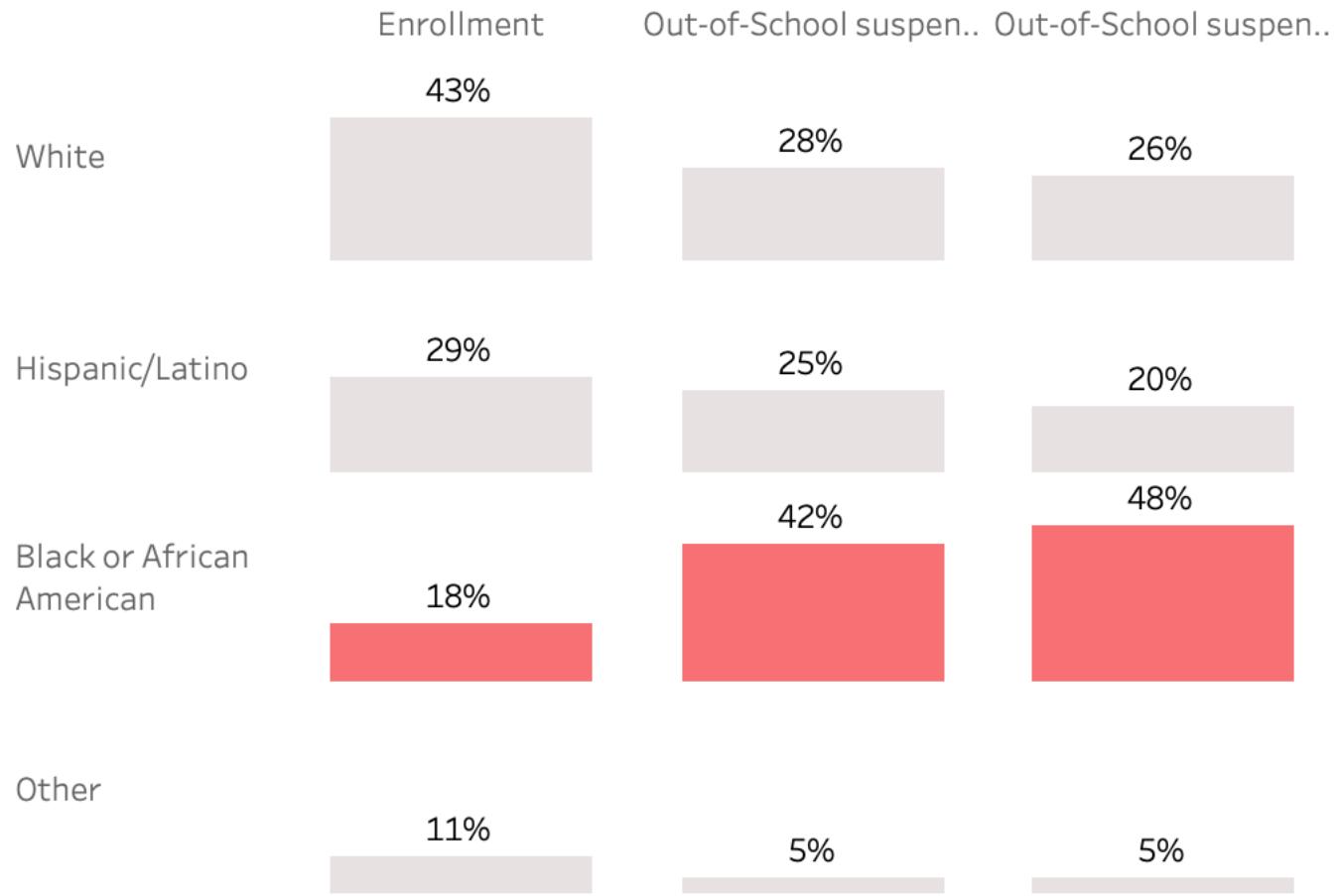


Race

- White
- Hispanic/Latino
- Black or African American
- Other

Шаг 3:
Группируем
малозначимые
группы, делаем
осмысленный
заголовок, выделяем
ярким ключевую идею

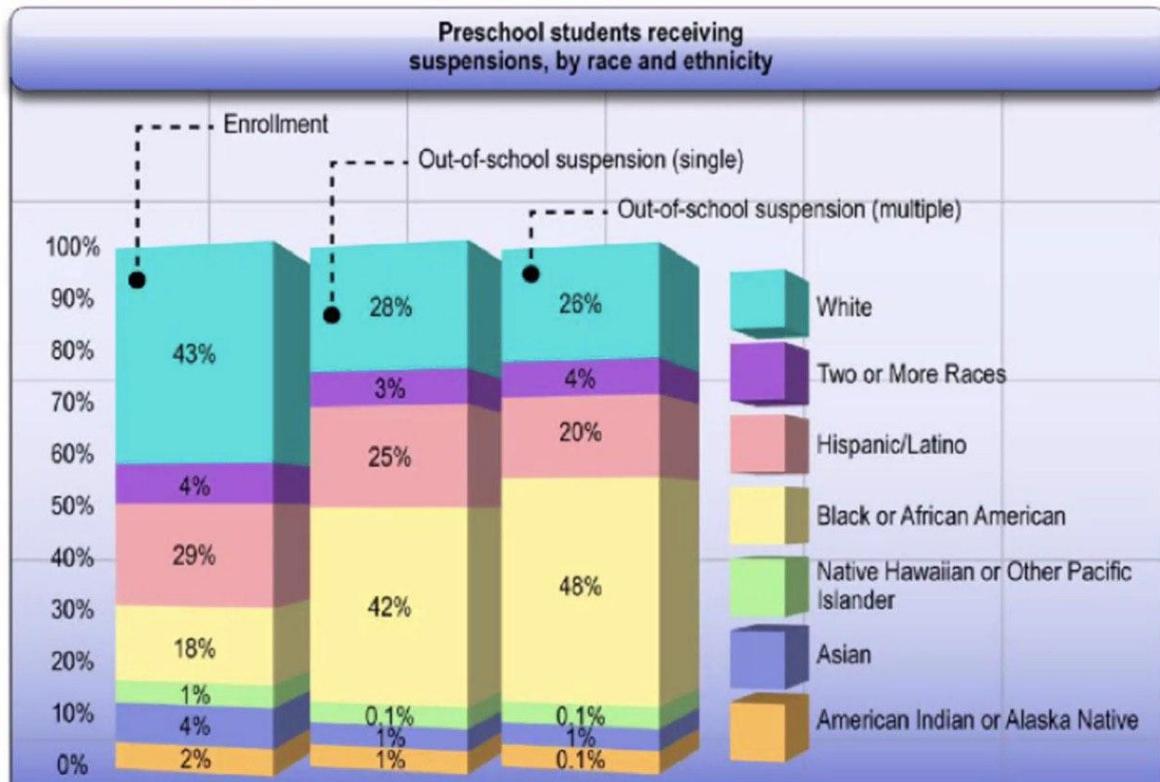
Black preschool students are more likely suspended than other children



Шаг 4:

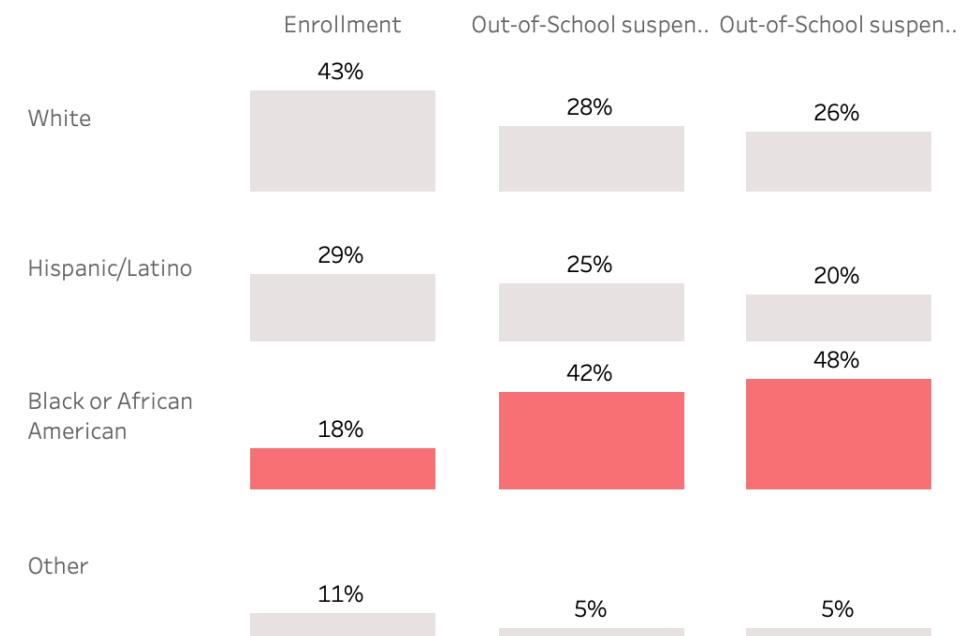
Выравниваем каждую категорию по горизонтали, уменьшаем количество цветов

До



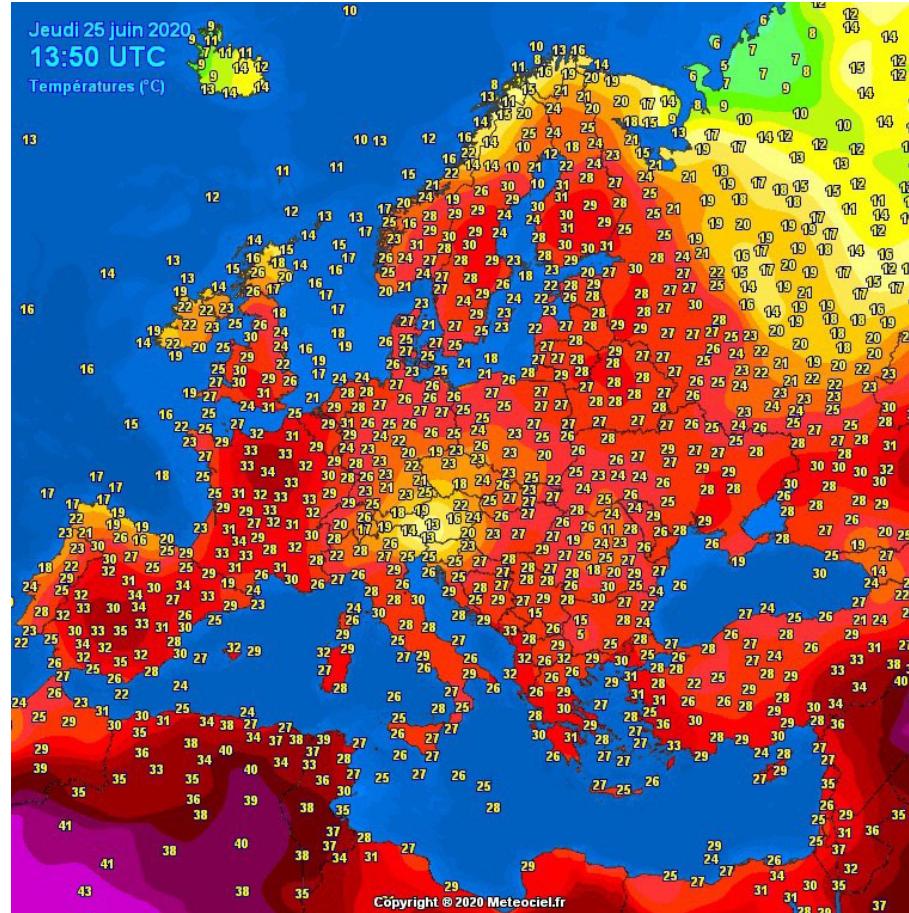
После

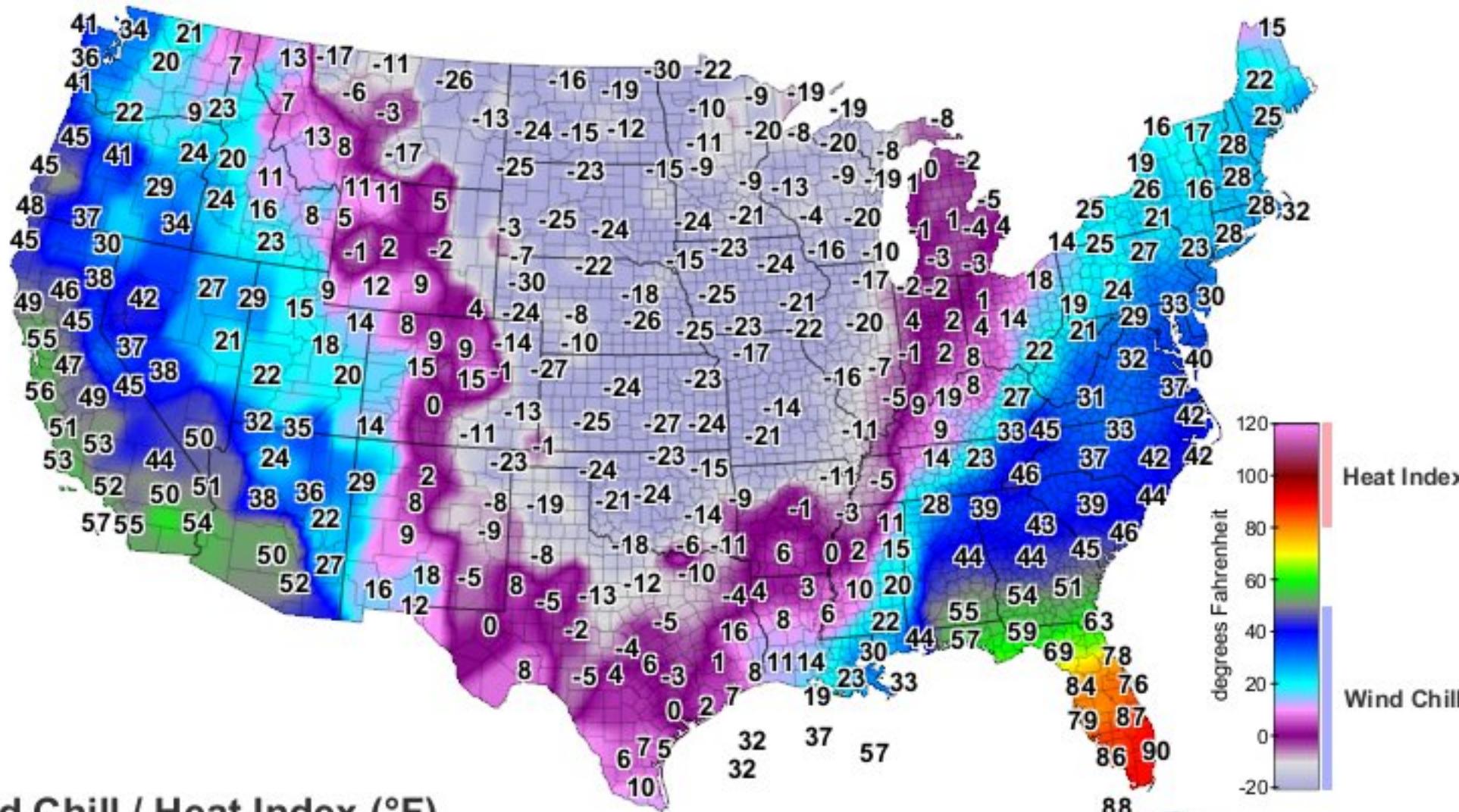
Black preschool students are more likely suspended than other children



**уберите всё,
что мешает**

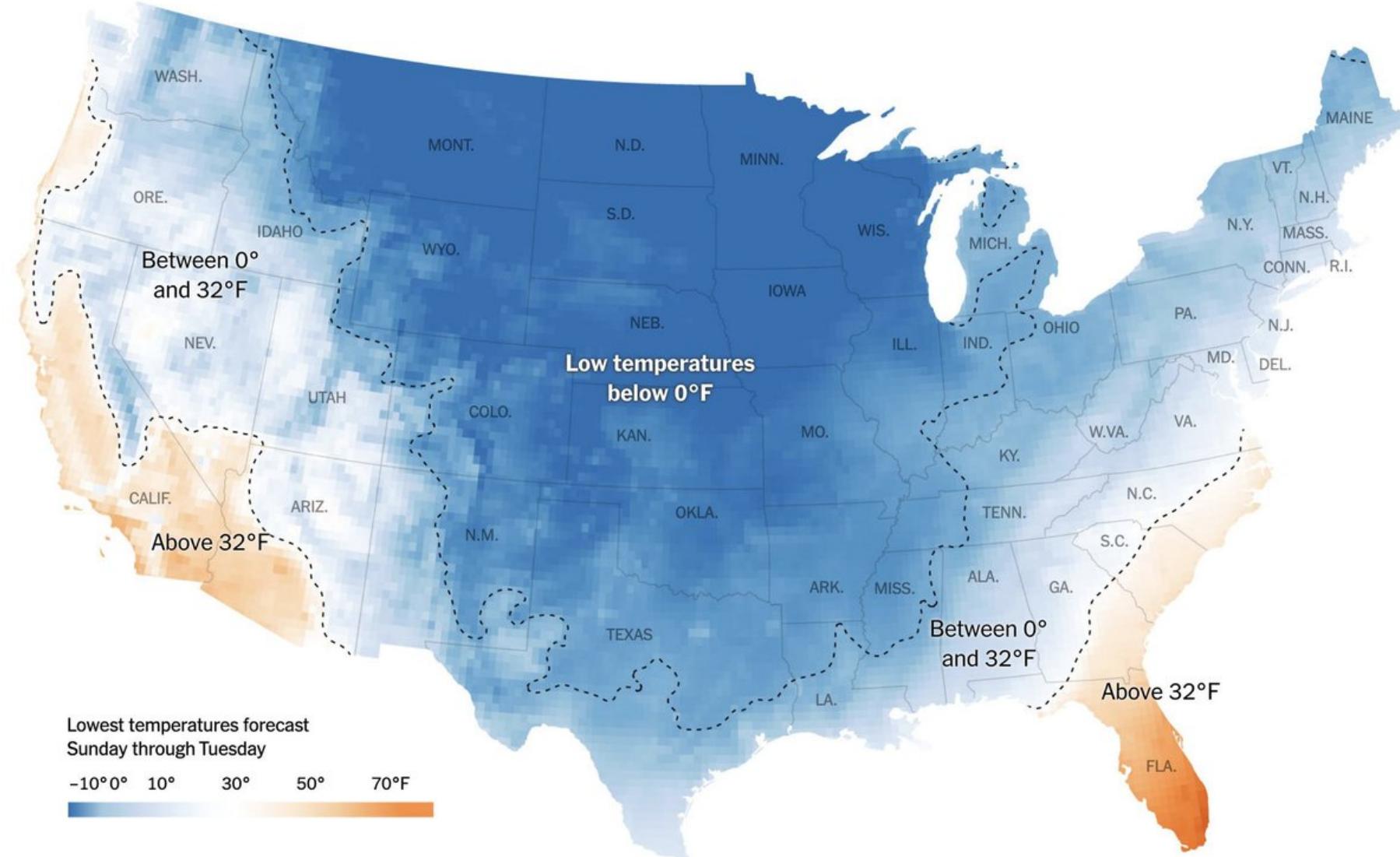
Слишком много подписей





Data provided by NOAA's National Weather Service. Created 11:10:34 AM February 15, 2021 CST. © Copyright 2021

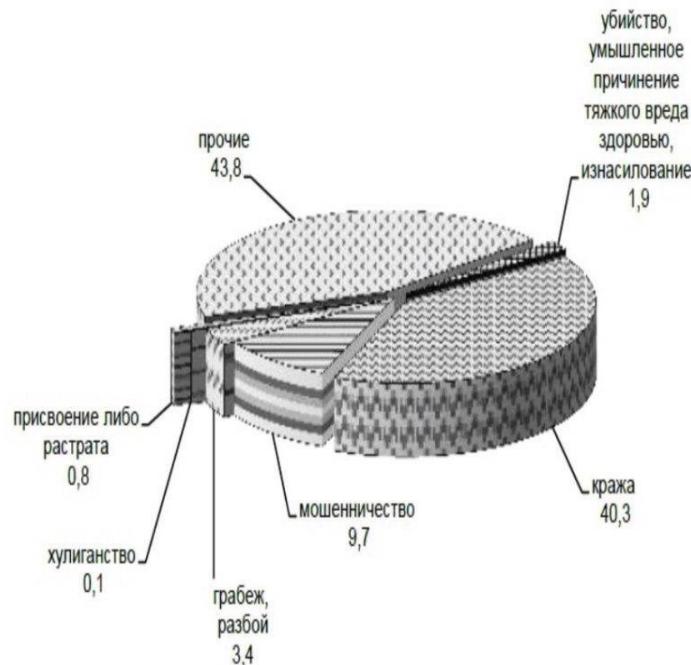




Трёхмерность

СТРУКТУРА ПРЕСТУПНОСТИ (в %)

январь-декабрь



**Найдите в данных
историю и добавьте
акценты**

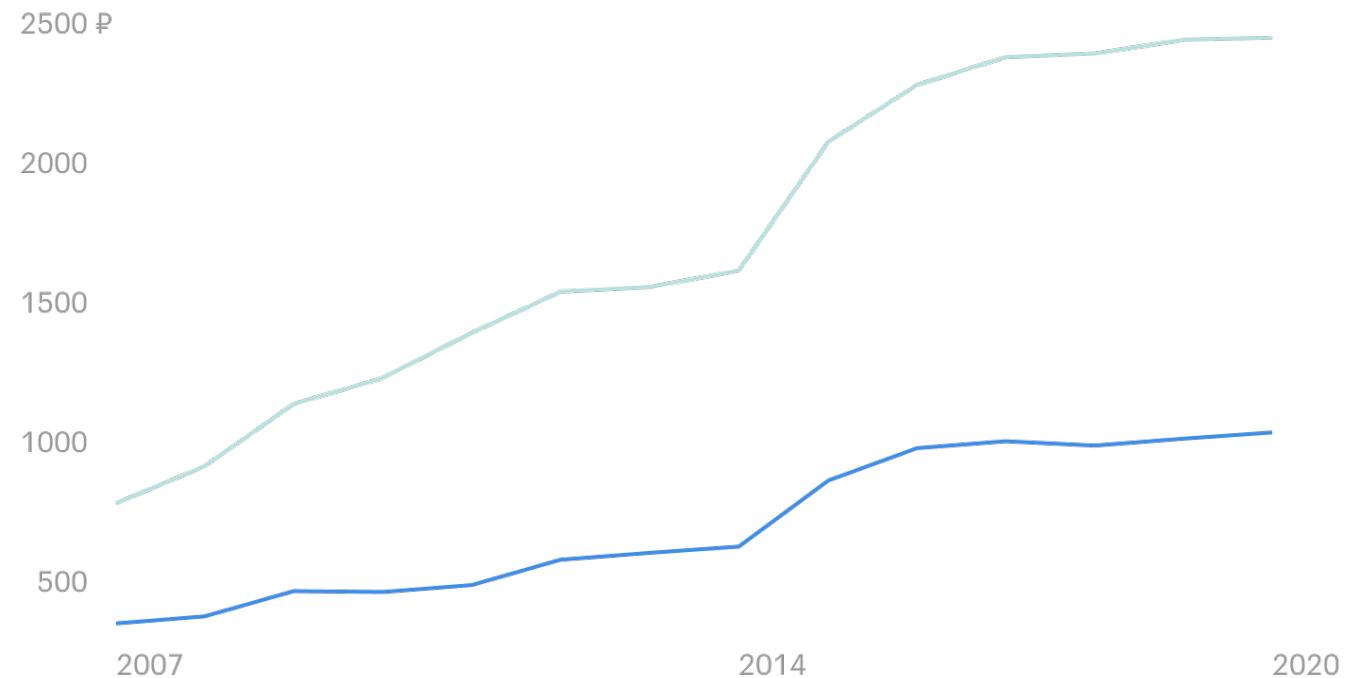
Заголовок

👎 **Не очень***:

Цены на растворимый и
зерновой кофе

✅ **Да:**

Цены на растворимый кофе
растут быстрее, чем на
зерновой



Заголовок



Не очень*:

15 самых распространенных
уличных топонимов на 2020
год



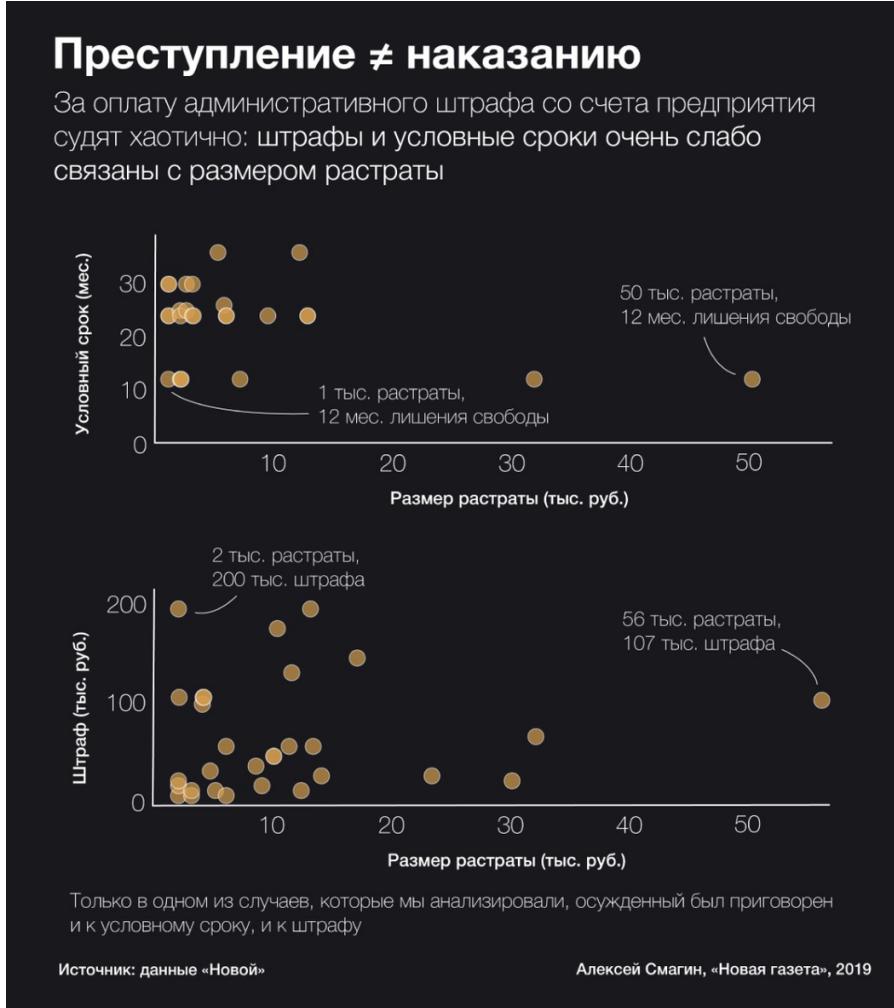
Да:

Улица Ленина на 14 месте
среди самых «популярных»
улиц
15 самых распространенных
уличных топонимов на 2020 год

Место	Улица	Количество
1	Центральная	25 860
2	Молодежная	17 727
3	Лесная	16 470
4	Школьная	15 913
5	Садовая	14 441
6	Новая	13 792
7	Советская	12 766
8	Набережная	11 065
9	Заречная	10 695
10	Полевая	10 542
11	Луговая	9 972
12	Зеленая	9 749
13	Мира	9 127
14	Ленина	7 652
15	Октябрьская	6 919

*окей, сторителлинговые заголовки не всегда уместны, но с ними часто лучше, чем без них

Заголовок + подзаголовок



Что писать? (опционально)

1. Что изображено на диаграмме (количество убийств в штате N, топ-10 компаний, которые производят колбасу)
2. Описание главной мысли графика
3. Афоризм для привлечения внимания (по аналогии с заголовками у статей)

Заголовок + подзаголовок

Где **алкогольная смертность** почти равна или больше
смертности от внешних причин

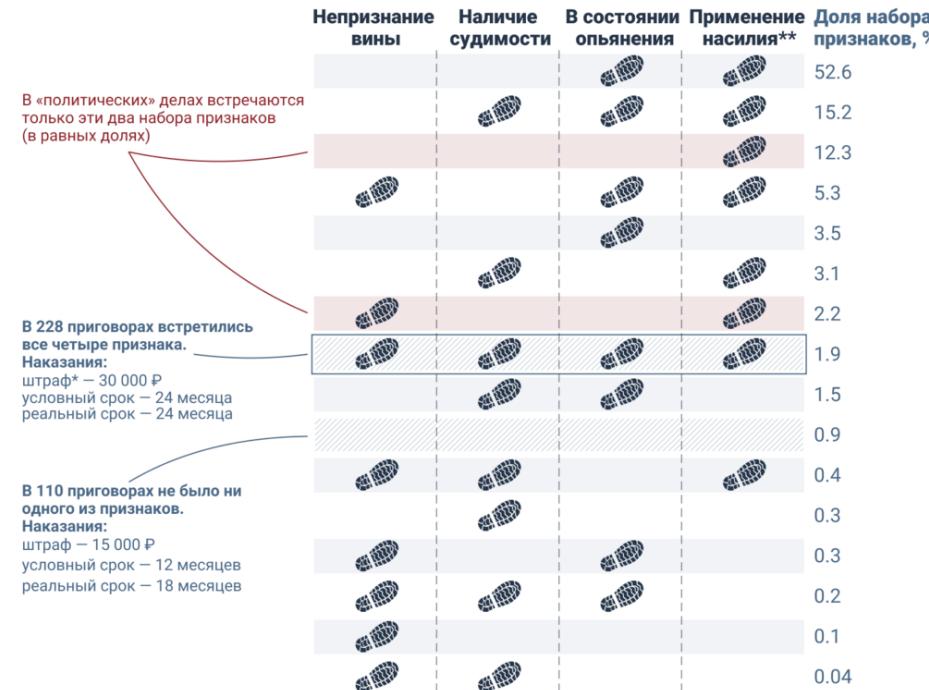
Человек, в 2018 году



Заголовок + подзаголовок

Судебное бинго

Какие наборы признаков чаще встречаются в приговорах по ч.1 ст.318 с 2016 по 2019 год (один набор – одна строка)



* здесь и далее указаны медианные значения

** согласно тексту приговора

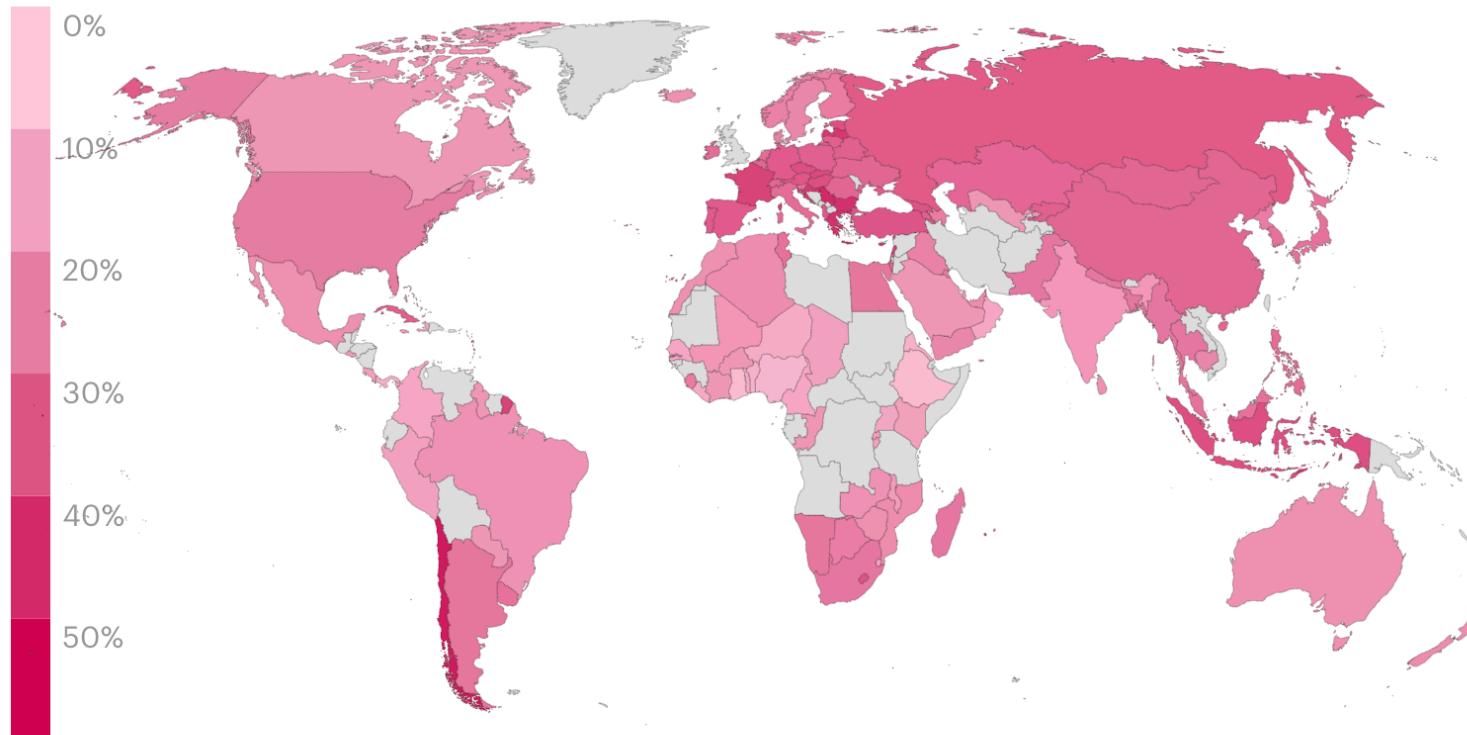
Источники: ГАС «Правосудие», Мосгорсуд, ОВД-Инфо, Медиазона

Артём Щенников, «Новая газета»

Заголовок + подзаголовок

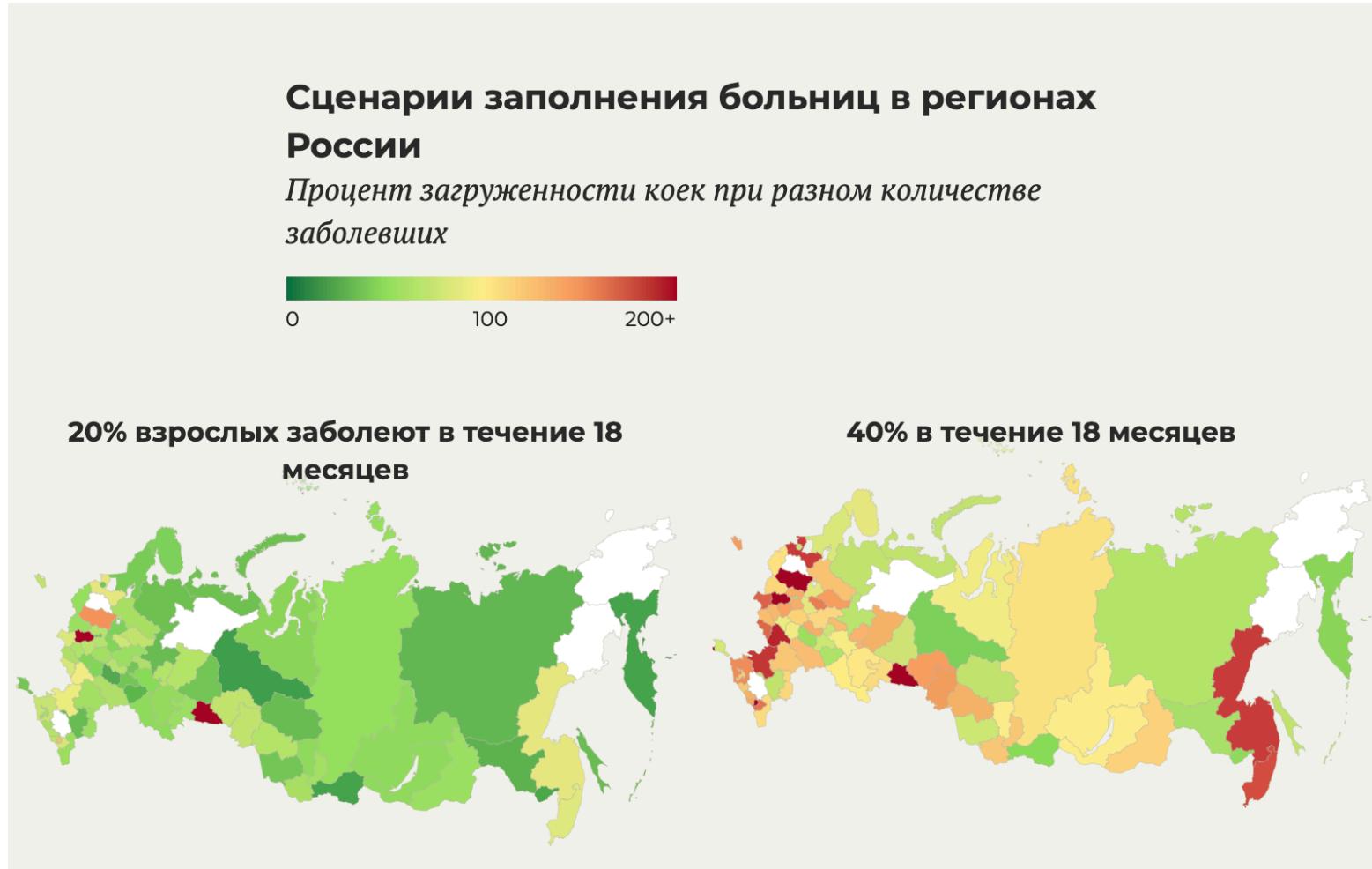
Доля курящего населения в странах мира в 2018 году

Среди граждан старше 15 лет



Наводите на страны, чтобы получить больше информации. Источник: [Всемирная организация здравоохранения](#)

Заголовок + подзаголовок



Цветовые акценты



Улица Ленина на 14 месте среди самых «популярных» улиц

15 самых распространенных уличных топонимов, 2020

Место	Улица	Количество
1	Центральная	25 860
2	Молодежная	17 727
3	Лесная	16 470
4	Школьная	15 913
5	Садовая	14 441
6	Новая	13 792
7	Советская	12 766
8	Набережная	11 065
9	Заречная	10 695
10	Полевая	10 542
11	Луговая	9 972
12	Зеленая	9 749
13	Мира	9 127
14	Ленина	7 652
15	Октябрьская	6 919



Улица Ленина на 14 месте среди самых «популярных» улиц

15 самых распространенных уличных топонимов, 2020

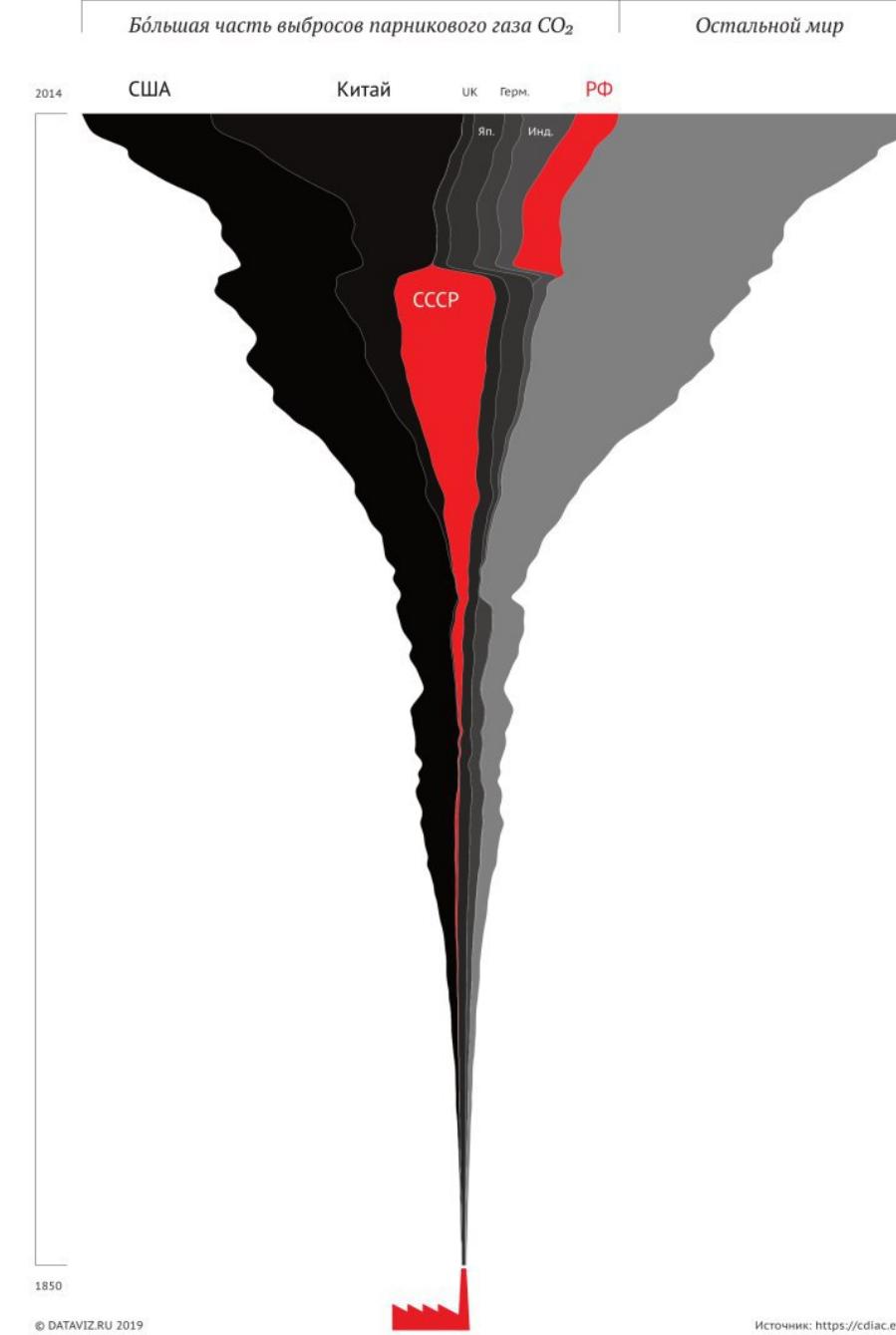
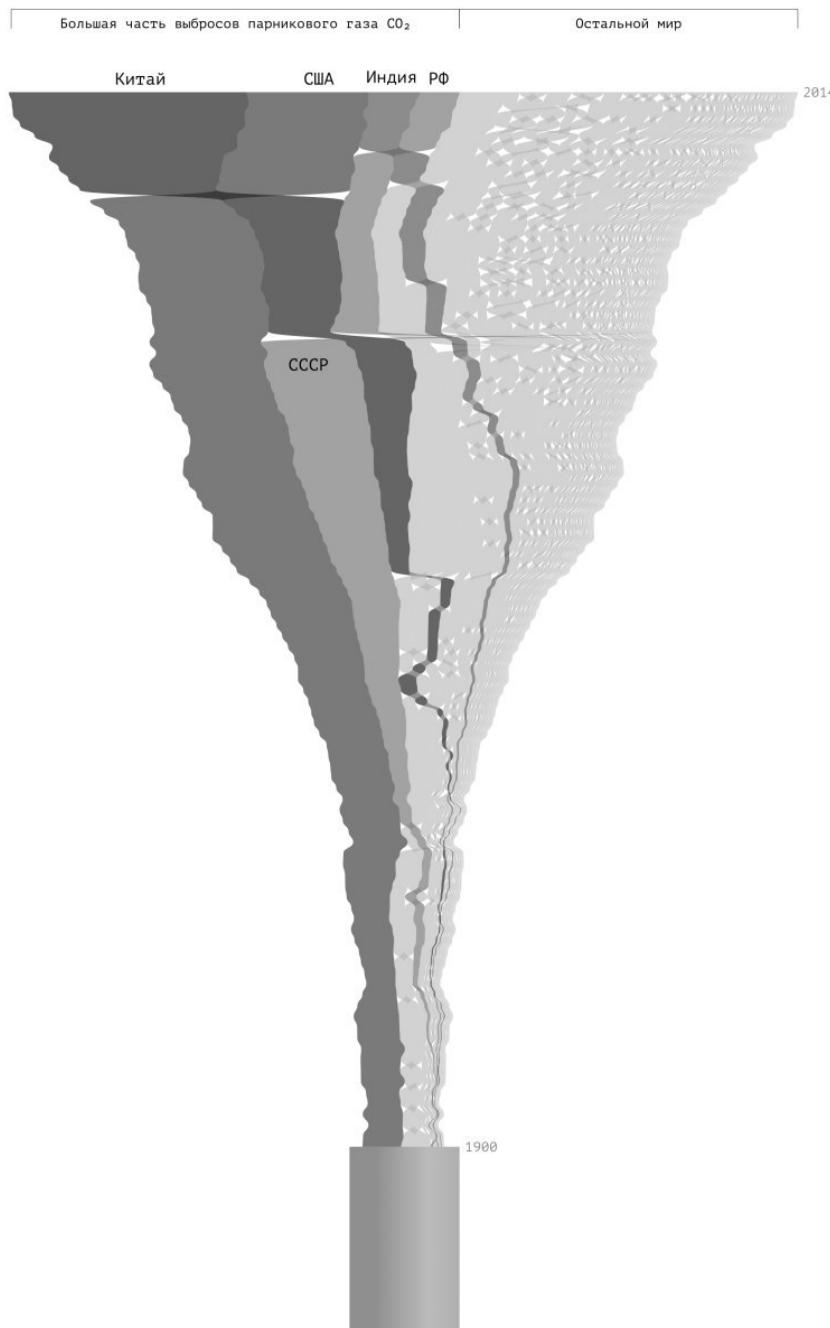
Место	Улица	Количество
1	Центральная	25 860
2	Молодежная	17 727
3	Лесная	16 470
4	Школьная	15 913
5	Садовая	14 441
6	Новая	13 792
7	Советская	12 766
8	Набережная	11 065
9	Заречная	10 695
10	Полевая	10 542
11	Луговая	9 972
12	Зеленая	9 749
13	Мира	9 127
14	Ленина	7 652
15	Октябрьская	6 919

Цветовые акценты



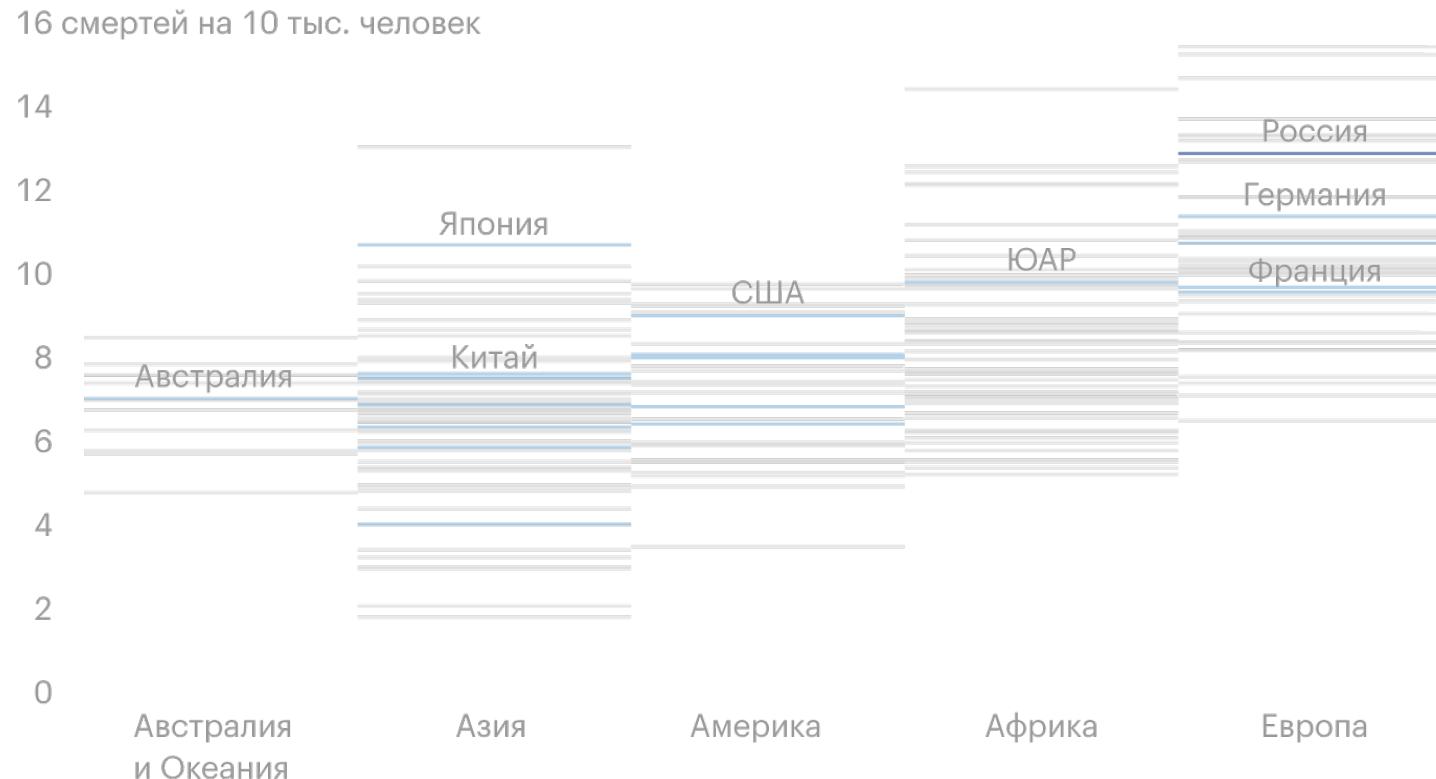
Кто виноват в изменении климата?

Кто украл снег на Новый год?



Цветовые акценты

Россия по уровню смертности обгоняет все страны из **G20**
и почти все страны Африки



% —

Japan, the “pioneer”

Japan has the most aged society in the world. In 2006, it became the first country with more than 20 percent of its population at 65 or older. By 2050, that is projected to exceed 35 percent.

Japan

% —

1960

1980

2000

2015

2040

2060

2080

Projection →

<https://graphics.reuters.com/JAPAN-AGING/010091PB2LH/index.html>

Симметрия в заголовке

В какие месяцы больше дней рождения?

Раньше дети чаще рождались в начале года.
В январе 1970 — на 4,9% больше, чем в другие месяцы этого года*

А сейчас,
почему-то, летом

Ну, число
рождений
в месяц явно
сезонно

* Известно сколько в России родилось людей в каждом месяце каждого (кроме 2003-2004) года. Считаем среднее за день по каждому месяцу. Потом отклонение этого числа от среднего за день по соответствующему году (с учетом високосных лет и числа дней в месяцах). Данные — «Демографические ежегодники» Росстата и приложения «Демоскопа».

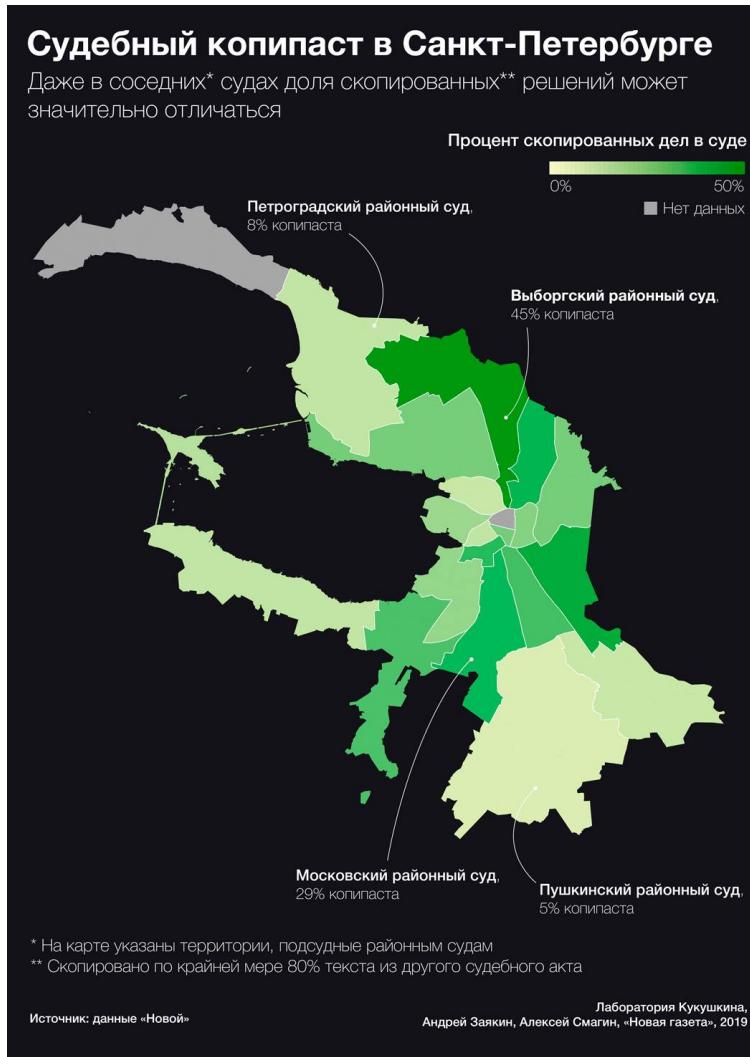


«Графики и жизнь», 9.20

Аннотации



Аннотации



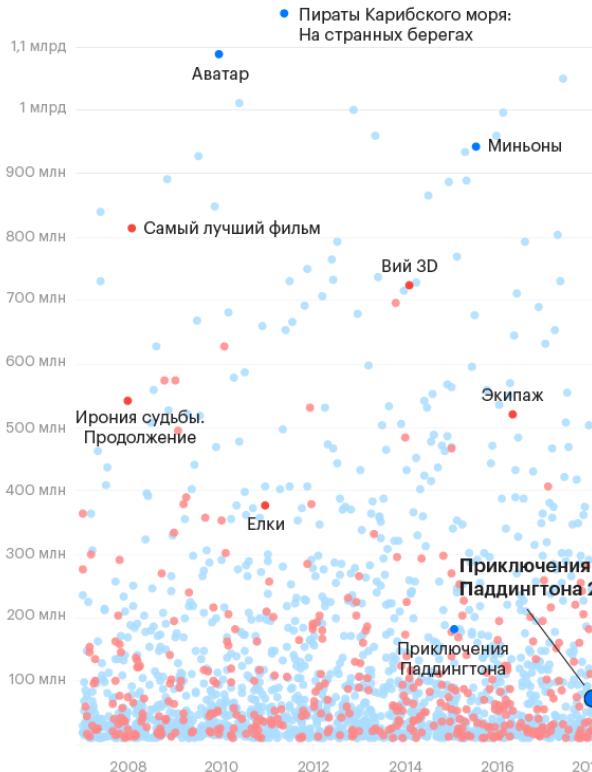
Аннотации

Старт проката

Кассовые сборы по итогам первых выходных всех фильмов, вышедших в прокат в России с 2007 года

● Российские ● Иностранные

₽



Учтены фильмы, сборы которых в первые выходные превысили ₽10 млн.
Суммы сборов указаны с учетом роста цен с момента выхода фильма.

Источник: данные РБК

© РБК, 2018

Аннотации



**Используйте
сортировку, чтобы
навести порядок**

Countries That Spend The Most On International Tourism \$ Billion

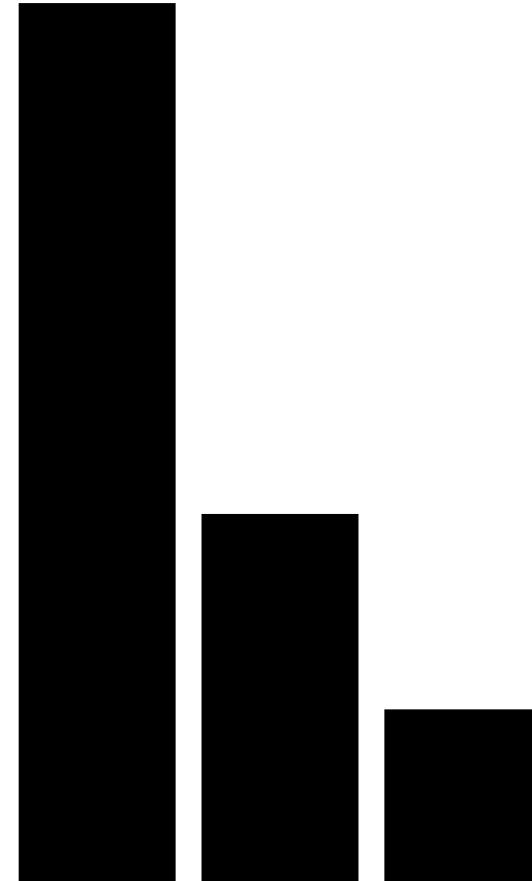


**Обязательно
указывайте
источник!**

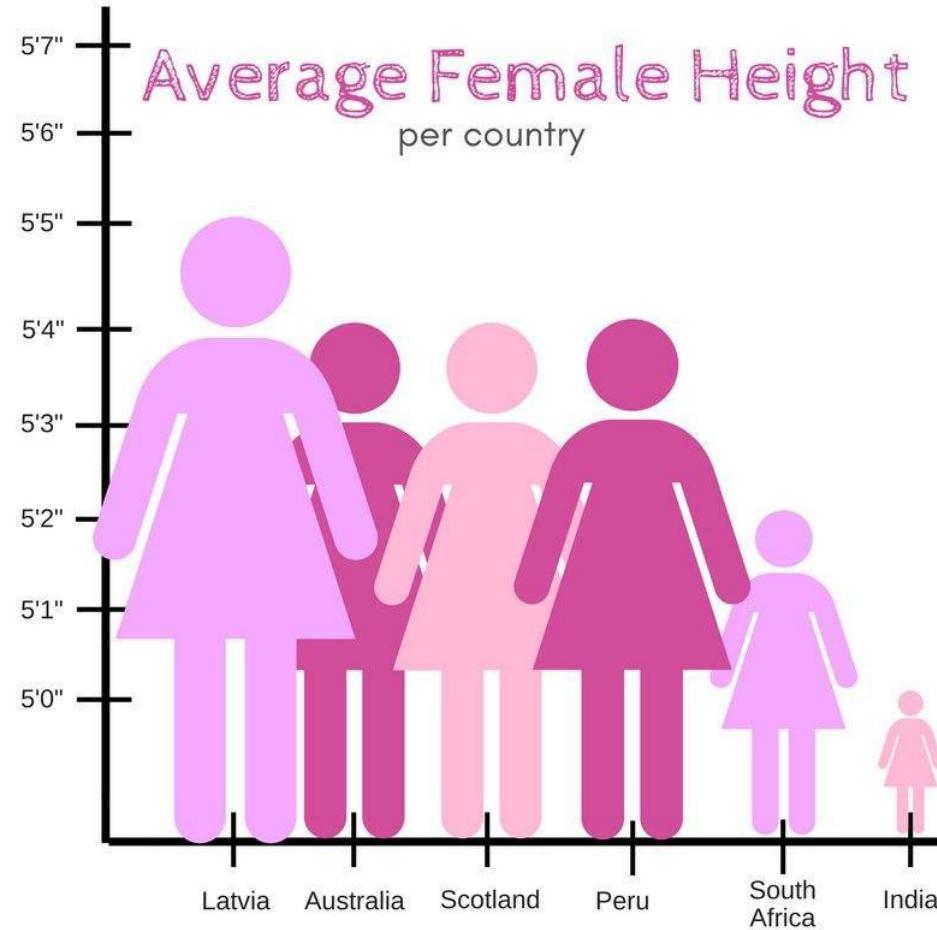
Барчарт

(или столбиковая, линейчатая диаграмма)

Лучший тип графика, когда надо показать, как величины соотносятся между собой

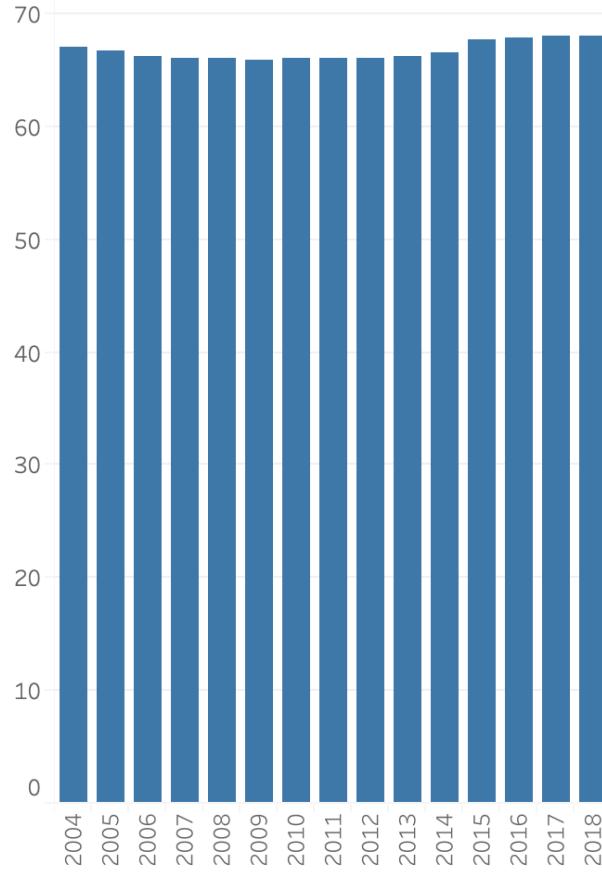


Всегда от нуля



Как показать изменение

Количество мужчин в РФ, млн. чел



Не очень:

Изменение величины мало по сравнению с самой величиной и плохо воспринимается

Как показать изменение

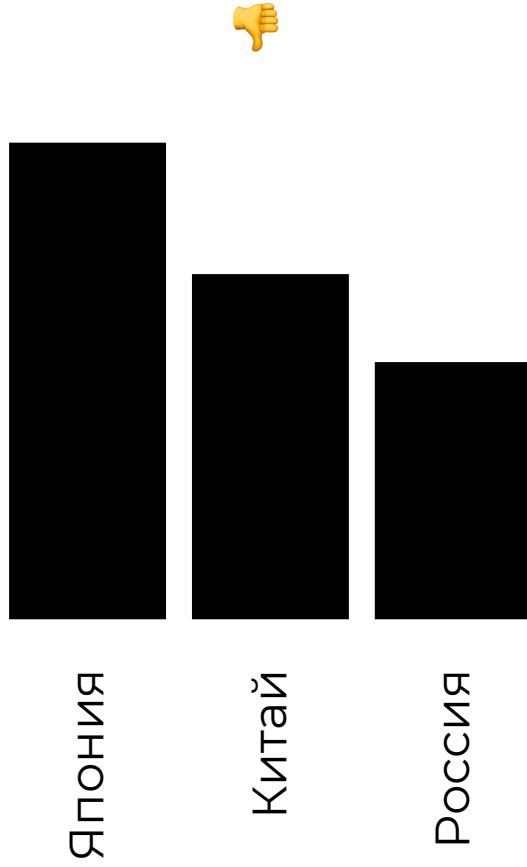
Как менялось количество мужчин по сравнению с прошлым годом, млн. чел



Да:

В высоту бара можно закодировать не саму величину, а её **изменение**

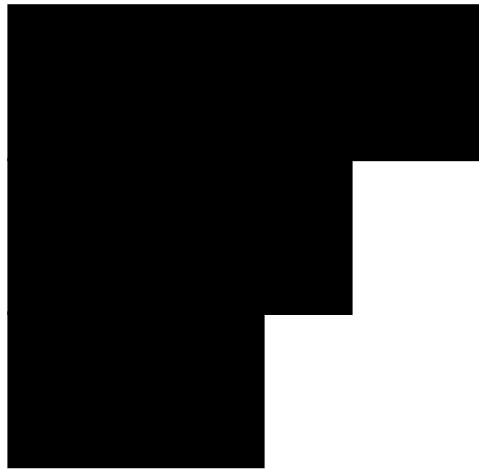
Если подписи большие



Бары не склеиваются



Япония



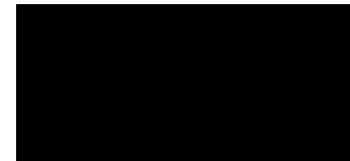
Китай



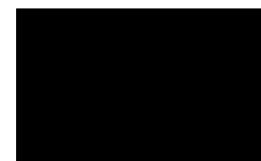
Япония



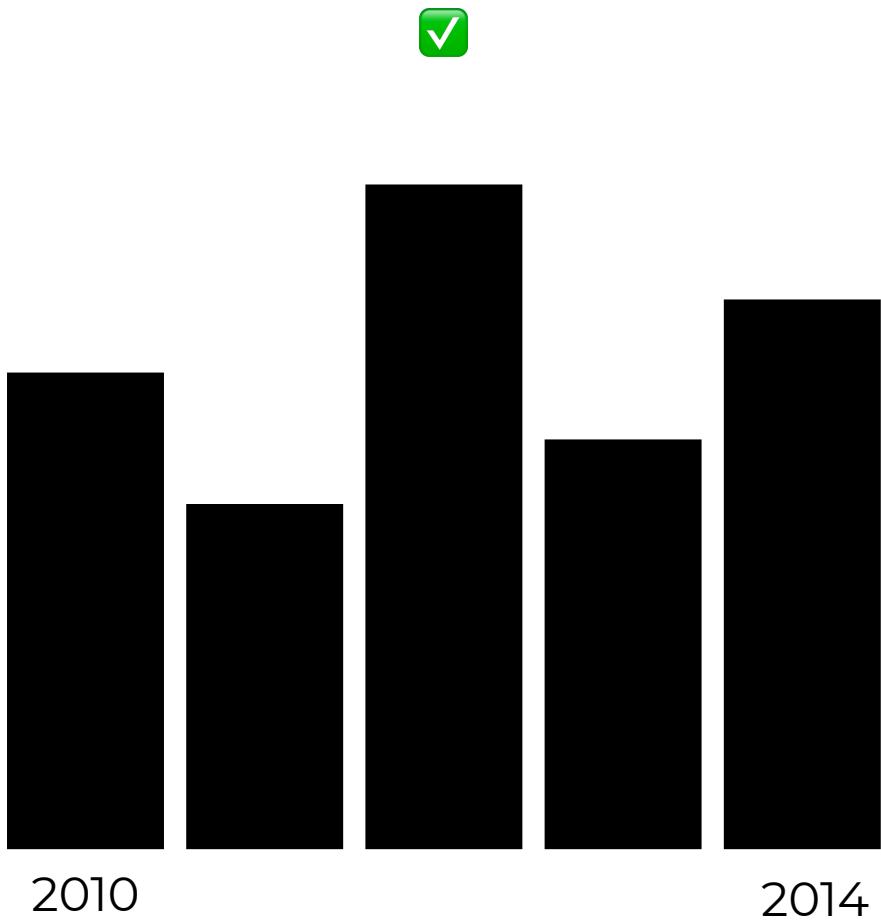
Китай



Россия

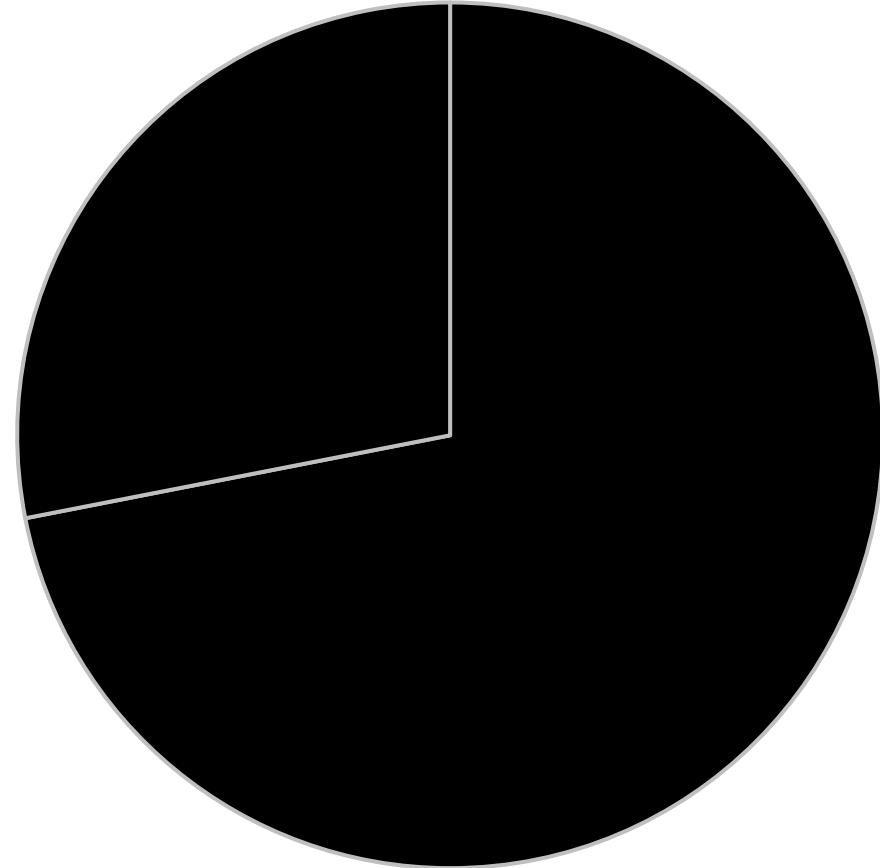


Если только это не время



Пайчарт

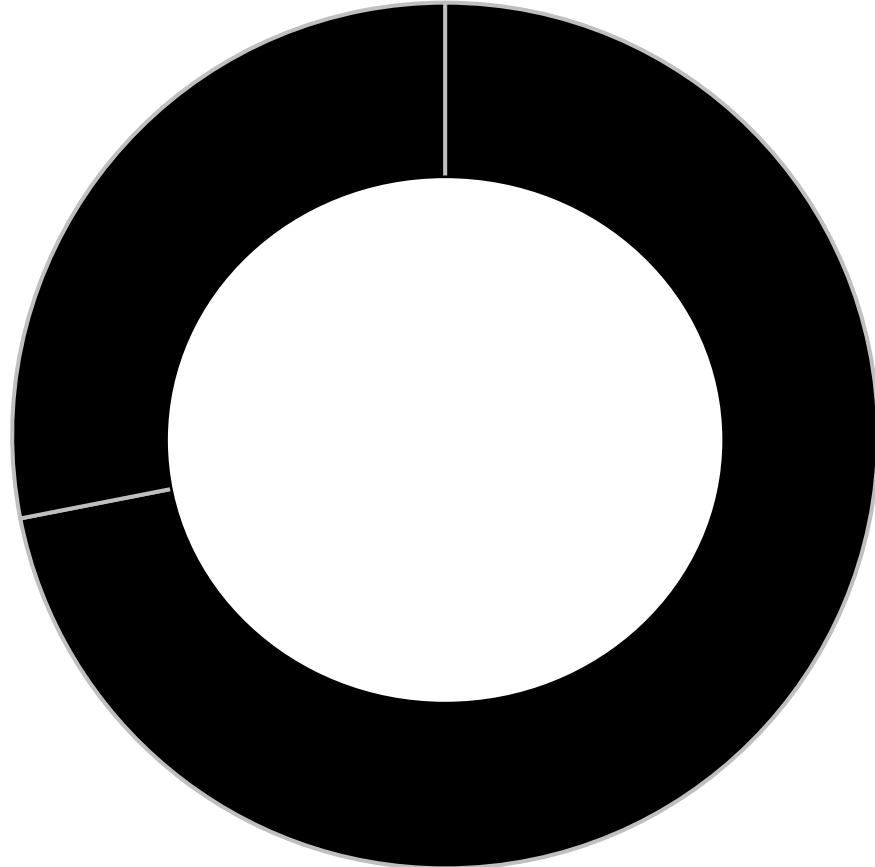
(или круговая диаграмма)



**Показывает какую часть от целого
составляет элемент**

Донат

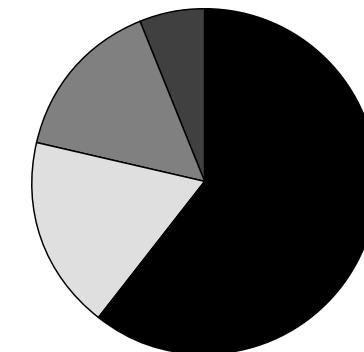
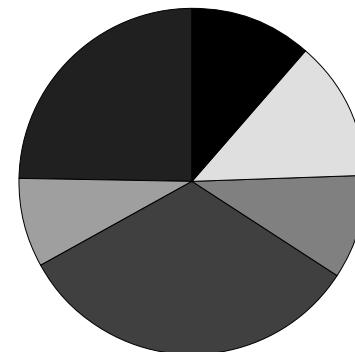
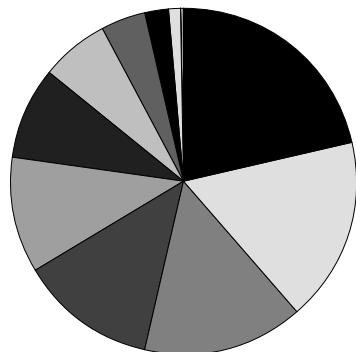
(или пончиковая диаграмма)



Пайчарт с дыркой

Правила пайчарта

1. Никому не говорить о пайчарте
2. Сектора пайчарта — это части целого и в сумме **всегда** должны давать 100%
3. Лучше не использовать больше 5-6 секторов (можно объединять мелкие сектора в «другое»)
4. Сектора должны идти по часовой стрелке от большего к меньшему. Сектор «другое» всегда последний.



Что можно объединить в пайчарт



Да

Национальный состав класса:
русские — 76%, украинцы — 15%,
татары — 6%, остальные — 2%



Нет

Доля людей, одобряющих
действия правительства: среди
лиц, имеющих высшее
образование — 15%, среди лиц,
имеющих среднеспециальное
образование — 30%, среди лиц
без образования — 55%

Результаты выборов президента
России: Путин — 77%, Грудинин —
12%, Жириновский — 6%,
остальные — 5%

Топ-3 кандидата на выборах
президента России: Путин — 77%,
Грудинин — 12%, Жириновский —
6%

Когда использовать пайчарт

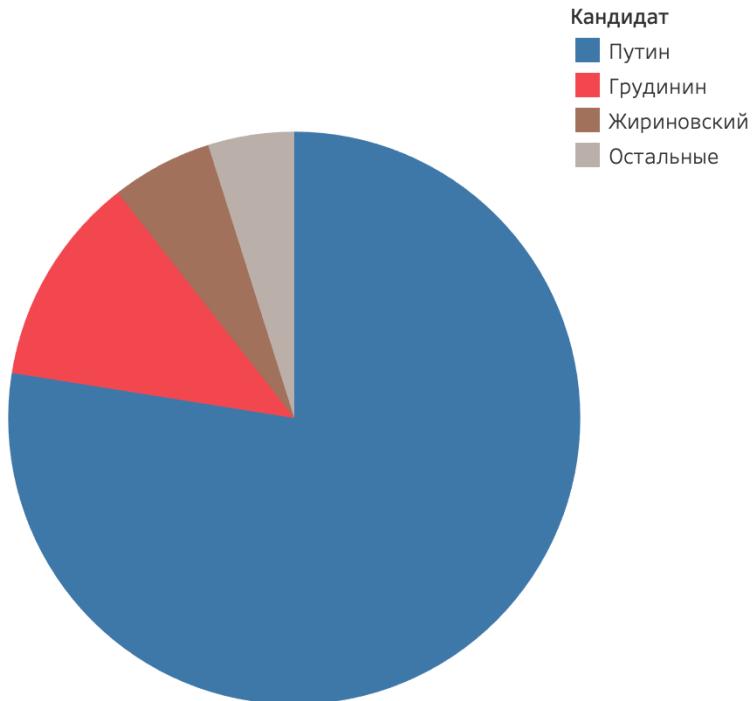
Сколько голосов набрали кандидаты в президенты
2018 года



Если нужно **сравнивать** доли голосов за разных кандидатов, лучше использовать бары

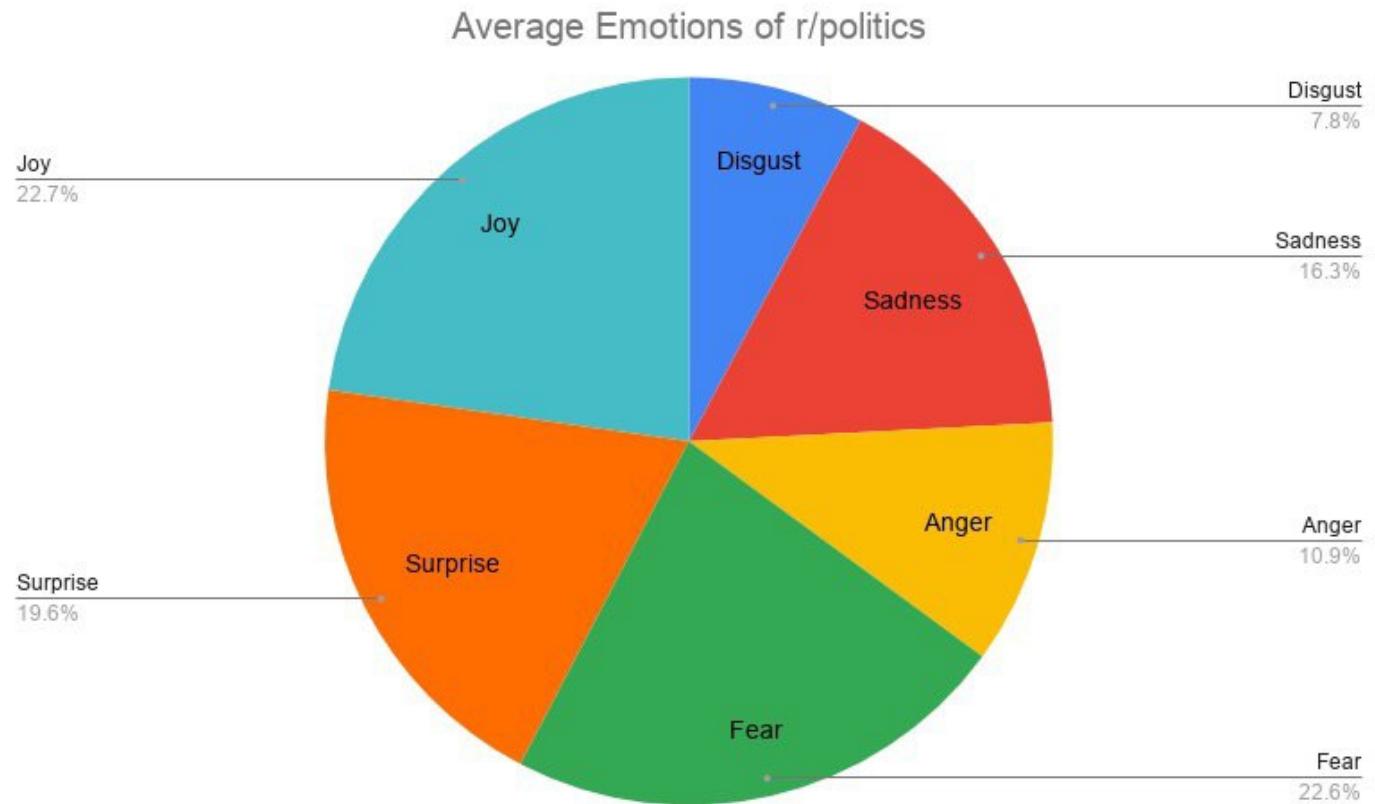
Когда использовать пайчарт

Путин набрал 76% голосов на выборах президента 2018 года



Если показать, что Путин набрал очень много голосов в общей доле — лучше пайчарт

Когда использовать пайчарт



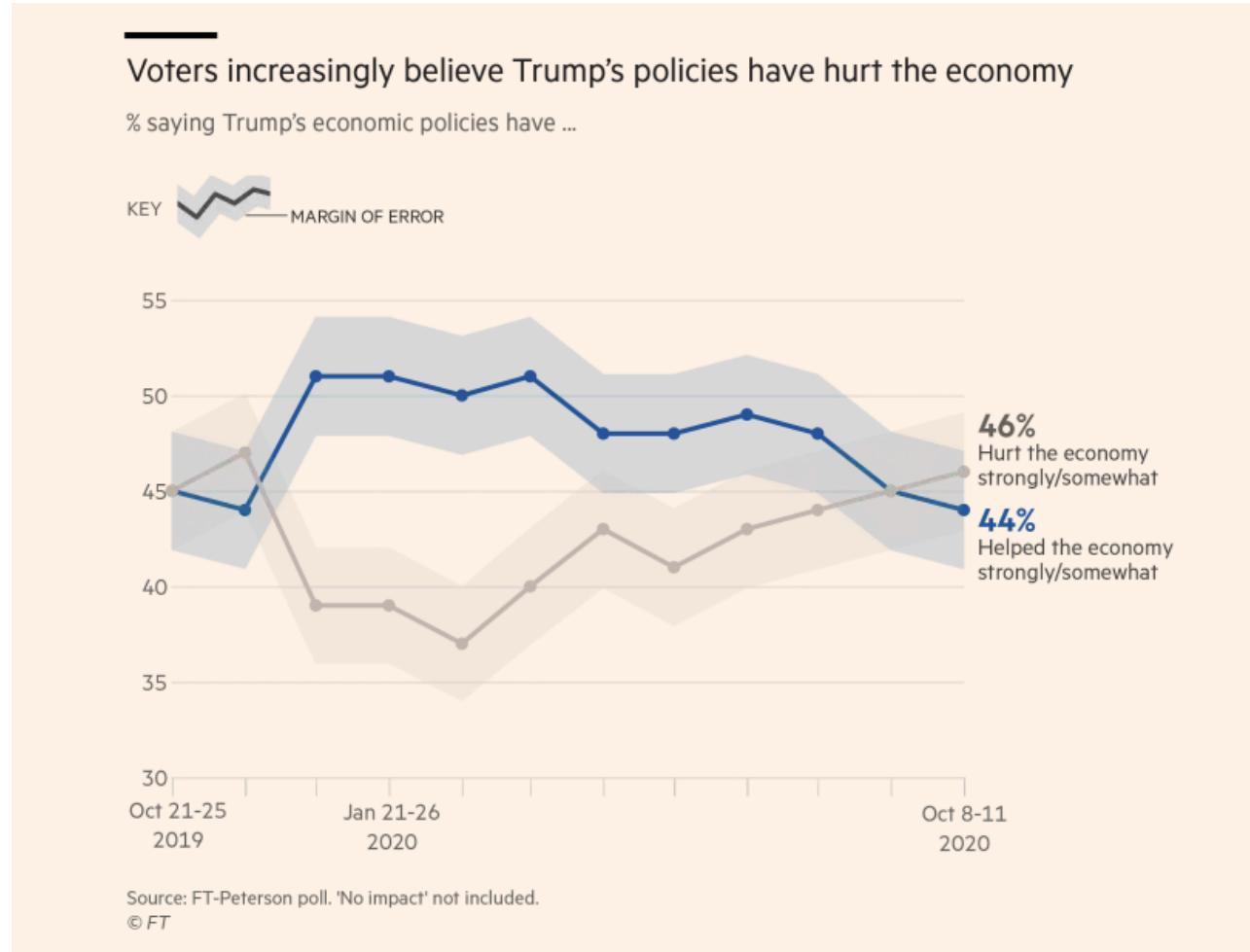
Помимо того, что тут нарушены все правила расположения секторов, пайчарт использовать бессмысленно, потому что сравнить величины крайне сложно

Линейный график



Показывает изменение величины во времени

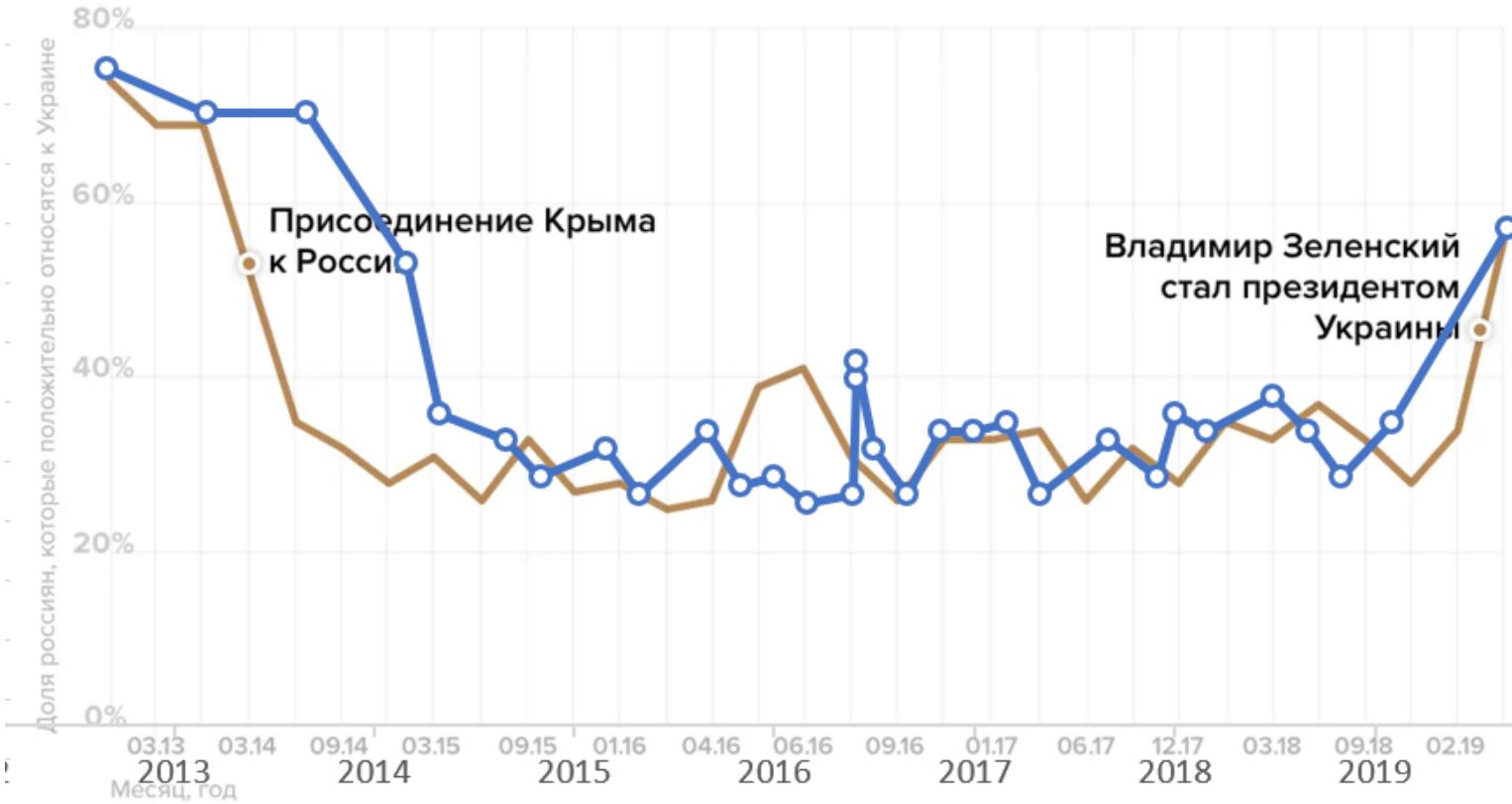
График можно строить не от нуля



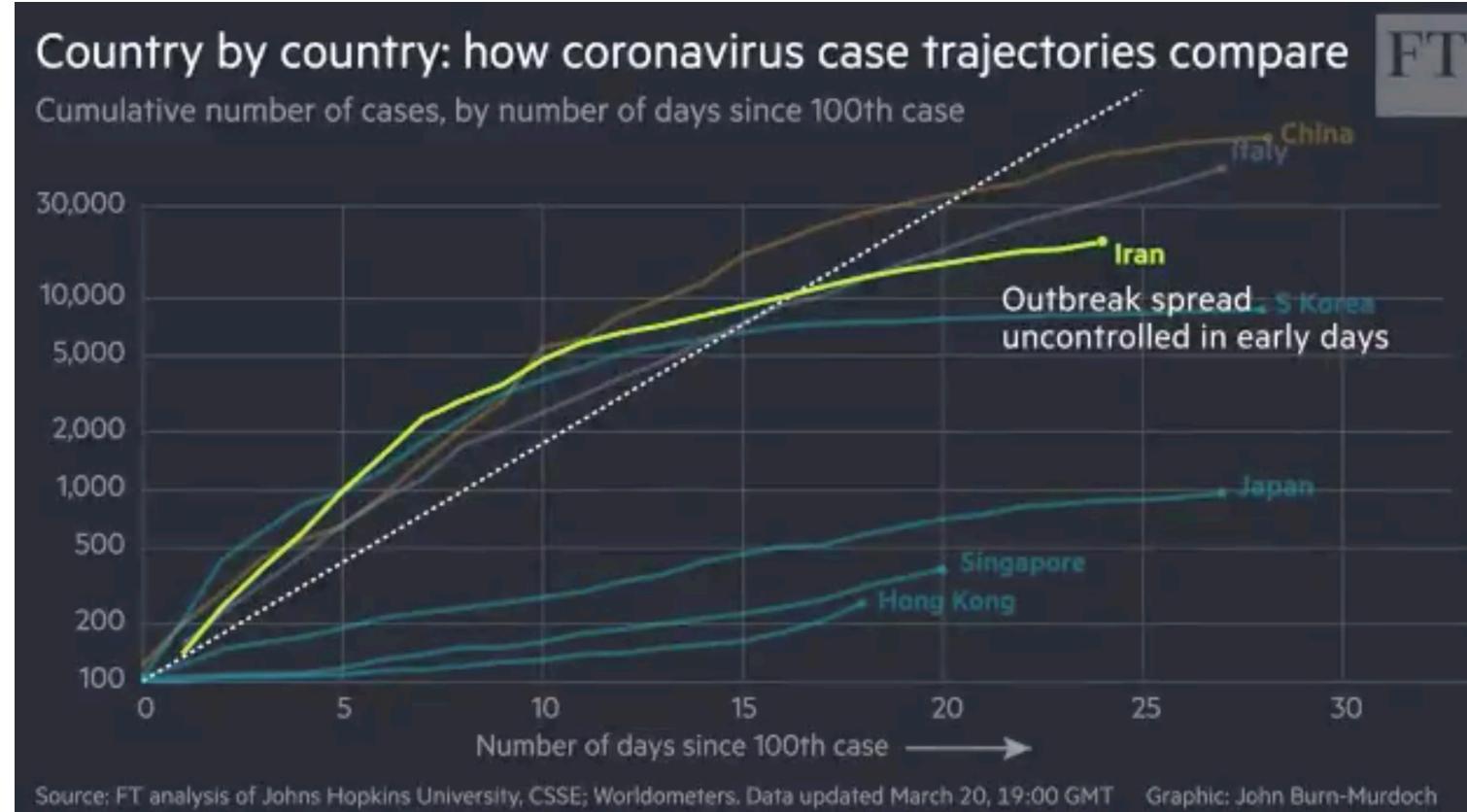
Что не так с этим графиком?



Ось времени нельзя деформировать



Логарифмическая шкала



[Интерактивная объяснялка](#)