Всемирная организация здравоохранения
Европейское региональное бюро

Инструменты для качественной визуализации данных: искусство использования диаграмм

Всемирная организация здравоохранения
Европейское региональное бюро

Инструменты для качественной визуализации данных: искусство использования диаграмм

Резюме
Визуализация данных – это набор методов, которые позволяют использовать визуальное представление для изучения, анализа и коммуникации количественных данных. Это помогает замечать в количественных данных тенденции и закономерности. Конечная цель визуализации данных – способствовать принятию более эффективных решений и мер. Чем больше становятся объемы доступных нам данных, тем важнее иметь возможность интерпретировать постоянно увеличивающиеся массивы информации, и визуализация данных позволяет эффективно решить эту задачу. Качественная визуализация данных – залог необходимого воздействия отчетов в области здравоохранения на целевую аудиторию. В настоящем документе даются практические советы по подготовке качественной визуализации данных, призванной повысить убедительность программных тезисов.

Настоящее руководство подготовлено в рамках работы Европейского регионального бюро ВОЗ по содействию государствам-членам в укреплении их информационных систем здравоохранения (ИСЗ). Оказание странам помощи в подготовке качественной информации по вопросам здравоохранения и создании институциональных механизмов для разработки политики с учетом фактических данных, традиционно относится к приоритетным направлениям работы ВОЗ и остается таковым в рамках Европейской программы работы на 2020–2025 гг.

Ключевые слова
DATA VISUALIZATION; DATA DISPLAY; COMMUNICATION; DECISION MAKING.

Номер документа: WHO/EURO:2021-1998-41753-58817

Содержание

Выражение признательности

Настоящий документ был разработан подразделением по вопросам данных, показателей и аналитики Отдела страновых стратегий и систем здравоохранения Европейского регионального бюро ВОЗ. Основным автором является Laurens Zwakhals. Marieke Verschuuren и David Novillo Ortiz руководили процессом подготовки доклада и предоставляли технические консультации при разработке концепции, написании и рецензировании документа. Особая благодарность выражается Natasha Azzopardi-Muscat за ее стратегическое руководство.

Для получения дополнительной информации свяжитесь с подразделением по вопросам данных, показателей и аналитики (euhiudata@who.int).

Цель рекомендаций

Настоящее руководство подготовлено в рамках работы Европейского регионального бюро ВОЗ по содействию государствам-членам в укреплении их информационных систем здравоохранения (ИСЗ). Оказание странам помощи в подготовке качественной информации по вопросам здравоохранения и создании институциональных механизмов для разработки политики с учетом фактических данных, традиционно относится к приоритетным направлениям работы ВОЗ и остается таковым в рамках Европейской программы работы на 2020–2025 годы[1]. Качественная визуализация данных – залог необходимого воздействия отчетов в области здравоохранения на целевую аудиторию. В настоящем документе даются практические советы по подготовке качественной визуализации данных, призванной повысить убедительность программных тезисов.

1 Европейская программа работы. В издании: ЕРБ ВОЗ [веб-сайт]. Копенгаген: Европейское региональное бюро ВОЗ; 2020 (https://www.euro.who.int/ru/health-topics/health-policy/european-programme-of-work/about-the-european-programme-of-work).

Об иллюстрациях

Все иллюстрации в этом документе созданы автором, кроме рис. 15. Рис. 15 распространяется по лицензии Creative Commons (CC-BY 4.0). Его можно бесплатно копировать и распространять на любом носителе и любом формате. Источники данных указаны под каждой иллюстрацией.

Введение: понятие визуализации данных и ее значение

Визуализация данных – это набор методов, которые позволяют использовать визуальное представление для изучения, анализа и коммуникации количественных данных *(1)*. Это помогает замечать в количественных данных тенденции и закономерности. Конечная цель визуализации данных – способствовать принятию более эффективных решений и мер.

Чем больше становятся объемы доступных нам данных, тем важнее иметь возможность интерпретировать постоянно увеличивающиеся массивы информации, и визуализация данных позволяет эффективно решить эту задачу. Сказанное актуально не только для специалистов по обработке данных и аналитиков: владение методами визуализации данных необходимо в любой профессии. С необходимостью визуализировать данные сталкивается любой, кто работает в сфере финансов, маркетинга, дизайна, мониторинга здоровья населения или в любой другой сфере. Это подтверждает важность визуализации данных *(2)*.

Данные можно визуализировать на различных этапах анализа и коммуникации. Во-первых, это очень полезно на этапе **исследования данных**. Данные можно легко анализировать, визуализируя их с помощью инструментов статистического анализа и электронных таблиц. С помощью визуализации можно выявлять различные отношения, изучать распределения и предпринимать сравнения. Инструменты бизнес-аналитики также чрезвычайно полезны для понимания содержащейся в данных информации. На этапе исследования данных не столь важно, какой тип диаграммы выбрать, какие пояснительные надписи разместить и как оформить иллюстративный материал: главное, чтобы визуализация позволяла аналитику получить новую информацию.

После того как удалось достаточно глубоко разобраться в наблюдаемых тенденциях и закономерностях, начинается следующий этап – продумывание истории, которую необходимо рассказать целевой аудитории. На этом этапе происходит **презентация данных**. Цель презентации данных (иногда называемой представлением данных) – ознакомить аудиторию с конечными результатами анализа. Они могут быть представлены в книге, отчете, презентации или на странице в интернете. Для того чтобы донести знание до широкого круга адресатов, необходимо представить его в четкой и понятной форме. Именно на этом этапе большое значение приобретают тип диаграммы, пояснительные надписи и визуальное оформление материала. Настоящий документ посвящен визуализации данных как раз на этом этапе – когда, переработав данные в информацию, мы должны помочь более широкой аудитории усвоить сделанные нами выводы. Иллюстрации должны говорить сами за себя.

Структура документа

- Изложение материала начинается с основ. В разделе 1 говорится о том, зачем нужна визуализация данных. На примере показывается, что визуализация данных помогает лучше понять результаты невизуальных видов анализа. В этом разделе также объясняется, как визуальные элементы помогают разобраться в отношениях, распределениях и сравнениях.
- В разделе 2 рассматриваются некоторые характеристики наборов данных. В зависимости от этих характеристик выбор тех или иных типов диаграмм может быть более уместным.
- В разделе 3 дается обзор наиболее часто используемых типов диаграмм. Вариантов существует бесконечное множество, но в этом разделе рассматриваются 13 основных типов, которые используются в 90% случаев.
- В разделе 4 даются рекомендации по составлению пояснительных надписей. Диаграмма состоит как из графических, так и из текстовых элементов; текстовые элементы позволяют поместить графическую иллюстрацию в контекст. Эти надписи важны для понимания того, что иллюстрирует диаграмма.
- В разделе 5 исследуются очевидные и не совсем очевидные подводные камни визуализации данных.
- Наконец, в разделе 6 обсуждаются некоторые принципы визуального оформления диаграмм. Единство оформления особенно важно для восприятия материала аудиторией в тех случаях, когда в документе используется больше одной диаграммы. После того как читатель привыкнет к определенному оформлению, ему будет легче интерпретировать следующую диаграмму, выполненную в таком же стиле.

1. Зачем нужна визуализация данных?

Визуализация необходима, поскольку выявить закономерности и тенденции, глядя на визуально обобщенную информацию гораздо проще, чем просмотрев тысячи строк в электронной таблице. Так устроен человеческий мозг. Учитывая, что цель анализа – получить практически значимую информацию, визуализированные данные гораздо ценнее. Даже если аналитик данных способен выявить закономерности и извлечь информацию из данных без визуализации, донести полученные результаты до других людей и объяснить их значение без диаграмм куда сложнее.

1.1 Сила визуализации: пример

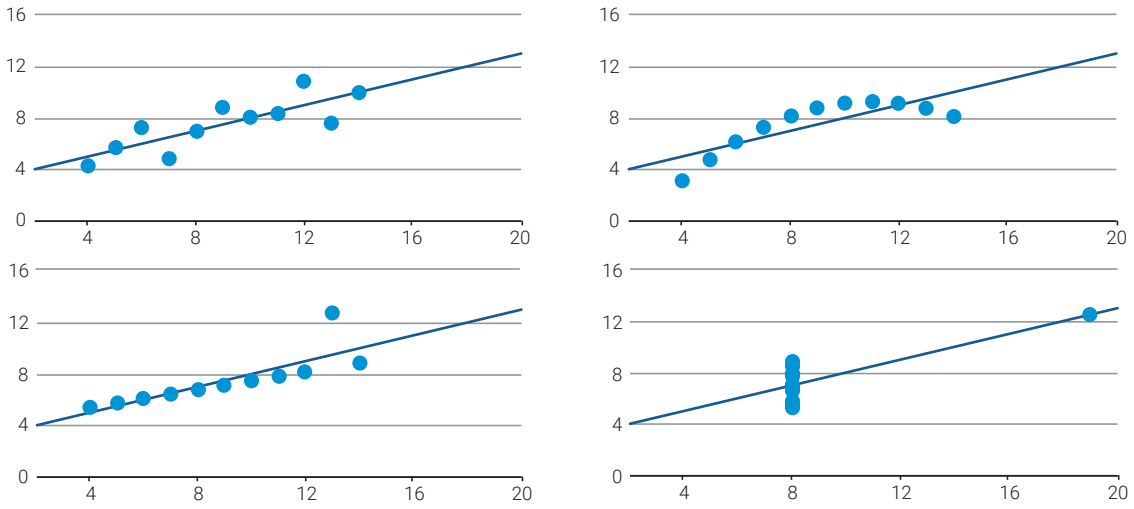
Силу визуализации можно проиллюстрировать на простом примере, известном как **квартет Энскомба** *(3)*. Это четыре набора данных, которые почти идентичны по описательным характеристикам, но имеют разное распределение и при графическом представлении дают совершенно разную картину. Каждый набор данных состоит из 11 точек (x, y). Эти наборы данных были разработаны в 1973 году специалистом по статистике Фрэнсисом Энскомбом, чтобы продемонстрировать, как важно перед анализом данных представить их в виде графика. Тем самым он опровергал расхожее представление в среде специалистов по статистике о том, что «числовые расчеты точны, а графики есть лишь грубое приближение».

В таблице 1 представлены наборы данных, разработанные Энскомбом. В первых трех из них значения x одинаковы. В конце таблицы приводится несколько описательных статистик. Основные статистические характеристики одинаковы, поэтому можно предположить, что четыре набора данных похожи, но, если представить их в виде диаграмм (рис. 1), различия становятся очевидны с первого взгляда. Визуализация данных может дать информацию, которую не всегда дают табличные данные и описательные статистики.

Таблица 1. **Квартет Энскомба**

	Набор данных I		Набор данных II		Набор данных III		Набор данных IV	
Наблюдение	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8,04	10	9,14	10	7,46	8	6,58
2	8	6,95	8	8,14	8	6,77	8	5,76
3	13	7,58	13	8,74	13	12,74	8	7,71
4	9	8,81	9	8,77	9	7,11	8	8,84
5	11	8,33	11	9,26	11	7,81	8	8,47
6	14	9,96	14	8,1	14	8,84	8	7,04
7	6	7,24	6	6,13	6	6,08	8	5,25
8	4	4,26	4	3,1	4	5,39	19	12,5
9	12	10,84	12	9,13	12	8,15	8	5,56
10	7	4,82	7	7,26	7	6,42	8	7,91
11	5	5,68	5	4,74	5	5,73	8	6,89
Сводная статистика								
Число	11	11	11	11	11	11	11	11
Среднее	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
Среднеква-дратичное отклонение	3,16	1,94	3,16	1,94	3,16	1,94	3,16	1,94
Коэффициент корреляции	0,82		0,82		0,82		0,82	

Рис. 1. **Визуализация квартета Энскомба**



Источник данных: Anscombe *(3).*

Без визуального представления информации аудитории может быть трудно понять истинное значение найденных закономерностей. Например, приводя таблицу со значениями ожидаемой продолжительности жизни населения за последние 20 лет, мы не даем аудитории веских причин проецировать эти данные на себя, однако если она увидит график, на котором виден рост или снижение ожидаемой продолжительности жизни – желательно в сравнении с данными других стран, – мы обязательно привлечем ее внимание.

1.2 Способы применения визуализации данных

Визуализация данных имеет множество применений, и каждый ее вид можно использовать по-разному. Подробно различные типы диаграмм рассматриваются в разделе 3; здесь же описаны лишь некоторые из наиболее распространенных способов использования визуализации данных.

1.2.1 Изменения с течением времени

Это, пожалуй, самый простой и распространенный способ применения визуализации данных, а значит, самый ценный. Этот способ наиболее распространен, поскольку в большинстве наборов данных присутствует аспект времени. Поэтому анализ данных часто начинается с выявления в них временных зависимостей. Для этого типа данных лучше всего подходят **линейные графики**.

Если необходимо отобразить много линий тренда сразу, линейные графики часто оказываются перегружены и образуют **диаграммы спагетти**. В качестве дополнительной графической переменной можно использовать цвет. Визуализация данных в данном случае столь же проста, как цветовое кодирование; такой прием может помочь при отображении больших наборов данных с большим количеством линий. Цветовая маркировка просто добавляет дополнительную переменную к координатам по осям X и Y. Получается **тепловая карта**.

1.2.2 Определение частот

Отображение частот – еще одно частое применение визуализации данных. Оно возможно, когда данные разделены на классы; классы могут быть порядковыми либо номинальными. Для этого типа визуализации лучше всего подходят **столбчатые диаграммы, линейчатые диаграммы** и **гистограммы**.

1.2.3 Определение отношений (корреляций)

Выявление корреляций – чрезвычайно ценное применение визуализации данных. Без визуализации трудно определить отношения между двумя переменными, однако знать о взаимосвязях в данных крайне важно. Для этого очень полезны **точечные диаграммы** и **пузырьковые диаграммы**.

Визуализация наборов данных Энскомба (рис. 1) является примером точечной диаграммы. Это отличный пример, показывающий ценность визуализации при анализе данных. Для понимания корреляций и выбросов одной описательной статистики недостаточно.

1.2.4 Изучение пространственных закономерностей

Если переменная имеет пространственный компонент, то хорошим способом визуализации распределения по области (например, по стране, региону или континенту) является **карта**. Этот способ может быть оптимален, например, для визуализации различий в уровне смертности между регионами страны.

Другая задача, для которой полезна карта, – визуализация риска пандемии через степень развития связей. Эпидемиологи используют этот тип визуализации для анализа того, как эпидемия распространяется по стране. Для этой цели создаются **карты потоков**.

1.2.5 Анализ ценностей и уверенности

Для определения сложных показателей, таких как значения с доверительными интервалами, необходимо учитывать множество различных переменных, что делает адекватный просмотр данных с помощью простой электронной таблицы почти невозможным. Для визуализации значений со степенью доверительной вероятности удобно использовать **диаграмму диапазонов с областями**.

1.2.6 Комбинации

Иногда может потребоваться сочетать разные способы визуализации данных. Если, например, необходимо визуализировать изменение частотного значения с течением времени, можно поместить рядом две линейчатых диаграммы, однако наилучшим вариантом будет **диаграмма наклона**. Сравнение частот с другой популяцией или целевым ориентиром может быть выполнено с помощью **диаграммы-шкалы**.

2. Характеристики данных

Любой набор данных имеет ряд характеристик. В этом разделе рассматриваются две характеристики, которые важно учитывать при выборе типа диаграммы: 1) содержит ли набор данных отдельные единицы данных или сгруппированные; 2) какая используется шкала измерения.

2.1 Индивидуальные или сгруппированные данные

Первой характеристикой набора данных является то, содержит ли он индивидуальные или сгруппированные данные. Так, индивидуальные данные могут быть визуализированы в виде **точечной диаграммы** или представлены в обобщенном виде как частоты на **гистограмме**.

Основная характеристика сгруппированных данных заключается в том, что отдельные наблюдения объединяются в группы. Примеры:

- количество новых случаев COVID-19 в сутки;
- количество людей в разбивке на классы по индексу массы тела;
- количество случаев смерти в разбивке по причине.

Сгруппированные данные лучше всего визуализировать с помощью **линейных графиков**, **столбчатых диаграмм** или **линейчатых диаграмм**.

2.2 Шкалы измерения

Вторая характеристика набора данных – это шкала измерения. В 1940-х гг. Стэнли Смит Стивенс выделил четыре шкалы измерения: **номинальную**, **порядковую**, **интервальную** и **шкалу отношений** *(4)*. Эти шкалы по сей день широко используются для описания характеристик переменной. Знание шкалы измерения переменной важно для выбора не только правильного метода статистического анализа, но и правильного типа диаграммы для визуализации данных.

2.2.1 Номинальная

Номинальная шкала характеризует переменную по принадлежности к одной из категорий, между которыми не предполагается какого-либо естественного порядка или ранжирования: это **качественная** шкала. Ее компоненты представляют собой связанные между собой отдельные группы, которые сами по себе не предполагают какого-либо строгого порядка; пример – список из 50 штатов Америки. Номинальные переменные можно кодировать числами, но порядок присвоения этих чисел будет произвольным, и любые вычисления с ними – будь то вычисление среднего, медианы или среднеквадратичного отклонения – будут бессмысленными.

Примеры номинальных переменных:

- группа крови;
- почтовый индекс;
- пол;
- раса;
- цвет глаз;
- политическая партия;
- религия;
- страны мира.

В целом для этой шкалы измерения лучше всего подходит **линейчатая диаграмма**. Поскольку качественные компоненты не имеют собственного порядка, их можно произвольно переставлять, чтобы выявлять закономерности в данных.

2.2.2 Порядковая

Порядковая шкала – это шкала, в которой важен **порядок следования** уровней, но не разница между значениями. Ее компоненты представляют собой элементы, которым свойственная некая естественная последовательность, например «холодно – тепло – горячо» или «белый – серый – черный».

Примеры порядковых переменных:

- социально-экономический статус (низкий уровень дохода, средний уровень дохода, высокий уровень дохода);
- уровень образования (начальное, среднее, степень бакалавра, магистра, доктора);
- доход (менее 50 000, 50 000–100 000, более 100 000);
- рейтинг удовлетворенности (совершенно не устраивает, не устраивает, нейтрально, устраивает, полностью устраивает).

Обратите внимание, что различия между смежными категориями не обязательно равнозначны. Например, разница между уровнями дохода «менее 50 000»

и «50 000–100 000» не равнозначна разнице между уровнями дохода «50 000–100 000» и «более 100 000».

В целом для этой шкалы измерения лучше всего подходит **столбчатая диаграмма**.

2.2.3 Интервальная

Интервальная шкала – это шкала, в которой уровни упорядочены, а интервалы между ними равны. Ее компоненты представляют собой элементы с постоянным числовым соотношением между собой (пример – последовательность минут).

Примеры интервальных переменных:

- температура (по Фаренгейту);
- температура (по Цельсию);
- кислотность (значение pH).

2.2.4 Шкала отношений

Переменная на шкале отношений имеет все свойства интервальной, но для нее также четко определен абсолютный ноль. Когда переменная равна 0,0, означаемая ею сущность отсутствует.

Примеры переменных, для которых используется шкала отношений:

- продолжительность;
- вес;
- длина;
- дозировка;
- температура в Кельвинах (0,0 К фактически означает *отсутствие тепла*);
- продолжительность выживания;
- пульс.

При работе с переменными на шкале отношений, в отличие от интервальных переменных, можно получить содержательную интерпретацию, оценив соотношение двух значений. Например, поскольку вес относится к шкале отношений, 4 г будет вдвое тяжелее, чем 2 г. Однако температуру 10 градусов по Цельсию не следует считать вдвое более высокой, чем 5 градусов по Цельсию. Если бы это было так, возникло бы противоречие, потому что 10 градусов по Цельсию – это 50 градусов по Фаренгейту, а 5 градусов по Цельсию – это 41 градус по Фаренгейту. Очевидно, что 50 – не вдвое больше, чем 41. Другой пример: среда с pH 3 не будет вдвое кислее, чем среда с pH 6, поскольку для pH не используется шкала отношений.

Как для интервальных переменных, так и для переменных на шкале отношений лучше всего использовать **линейные графики**.

3. Типы диаграмм

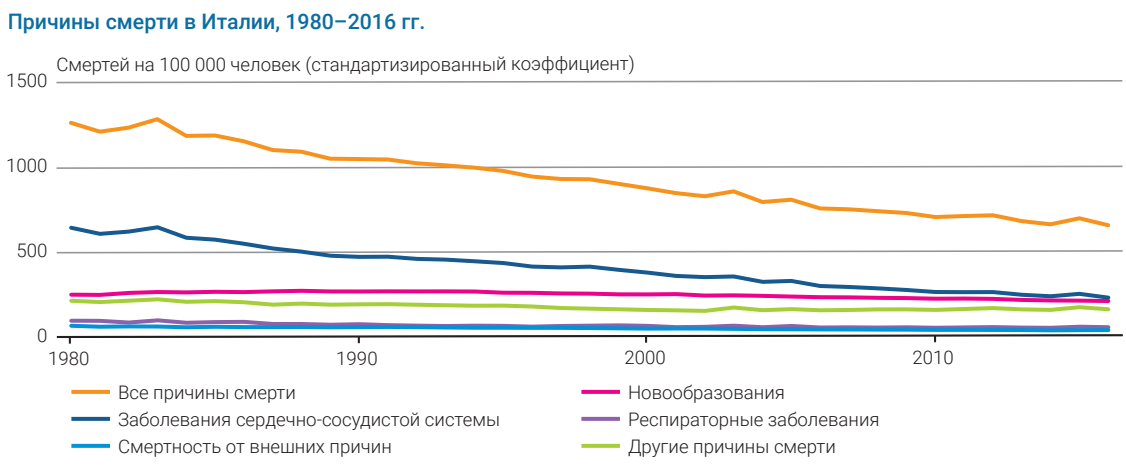
В предыдущих разделах говорится о том, как можно использовать визуализацию данных, а также изложены некоторые характеристики данных. В разделе 3 мы переходим к применению различных типов визуализации данных. Для визуализации данных существует множество инструментов. Некоторые из них в большей степени рассчитаны на ручную работу, а некоторые автоматизированы, но все они могут помочь в подготовке любого из типов визуализации, рассматриваемых в этом разделе.

В этом разделе рассматривается 13 различных типов диаграмм. В принципе их существует гораздо больше («спидометры», древовидные карты или иерархические схемы, столбчатые диаграммы с диапазоном, пузырьковые диаграммы без координат, ящичные диаграммы, диаграммы Санкея и т. д.), однако типы диаграмм, рассмотренные ниже, покрывают более 90% потребностей в визуализации данных.

3.1 Линейный график

Линейный график используется для отображения интервальной шкалы или шкалы отношений по оси X и количественного показателя по оси Y. Часто по оси X расположена шкала времени, но могут использоваться и другие непрерывные шкалы. Линейный график очень полезен для иллюстрации динамики количества смертей в разбивке по причинам, как показано ниже на рис. 2.

Рис. 2. Пример линейного графика

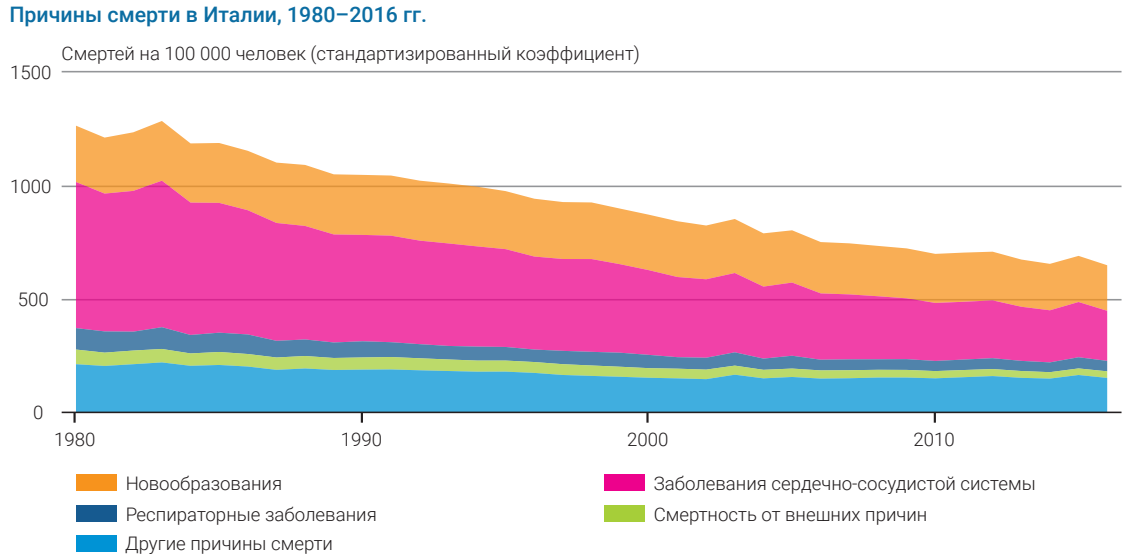


Источник данных: OECD.stat [онлайновая база данных]. Paris: Organisation for Economic Co-operation and Development; 2020 (https://stats.oecd.org/, по состоянию на 5 февраля 2021 г.).

3.2 Диаграмма с областями

Диаграмма с областями – это вариант линейного графика, где область под линией закрашена, чтобы подчеркнуть ее значимость. Если на графике отображено более одной переменной, диаграмму с областями необходимо использовать как **линейный график с накоплением**, как показано на рис. 3; значения отдельных категорий, показанных на диаграмме, складываются в общую сумму. На диаграммах этого типа нет необходимости отображать строку «Все причины смерти», так как отдельные количества сами складываются в итоговые суммы.

Рис. 3. Пример диаграммы с областями



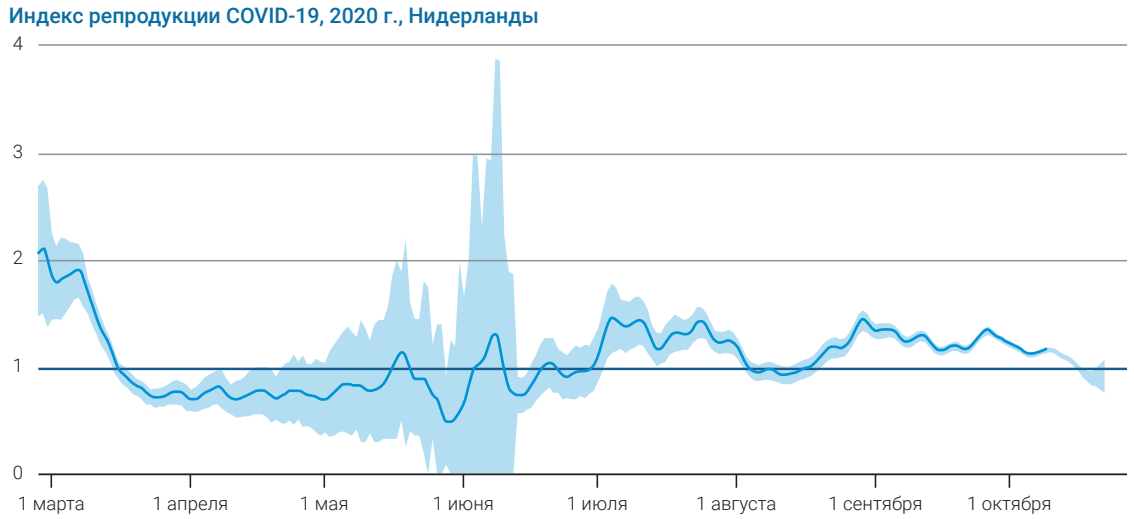
Источник данных: OECD.stat [онлайновая база данных]. Paris: Organisation for Economic Co-operation and Development; 2020 (https://stats.oecd.org/, по состоянию на 5 февраля 2021 г.).

3.3 Диаграмма диапазонов с областями

Диаграмма диапазонов с областями – это диаграмма с областями, где область определяется двумя значениями: верхним и нижним. Закрашивается только область между нижним и верхним значениями. Диаграмма диапазонов с областями в основном используется для отображения предполагаемого диапазона конкретного показателя.

Диаграмма диапазонов с областями часто сочетается с **линейным графиком**, показывающим средний уровень между верхним и нижним значениями (см. рис. 4). В этом примере показан предполагаемый индекс репродукции COVID-19 в Нидерландах с интервалом значимости, отображенным в форме диапазона с областями.

Рис. 4. Пример диаграммы диапазонов с областями в сочетании с линейным графиком



Источник данных: Covid-19 reproductiegetal [онлайновая база данных] [на голландском языке]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2020 (https://data.rivm.nl/geonetwork/srv/dut/catalog.search#/metadata/ed0699d1-c9d5-4436-8517-27eb993eab6e, по состоянию на 5 февраля 2021 г.).

3.4 Столбчатая диаграмма

Столбчатая диаграмма – лучший способ отобразить распределение значений порядковых переменных. На рис. 5 – пример распределения расходов граждан на здравоохранение по возрастным категориям.

Рис. 5. Пример столбчатой диаграммы

Источник данных: Kosten van ziekten 2015 [онлайновая база данных] [на голландском языке]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2019 (https://statline.rivm.nl/#/RIVM/nl/dataset/50040NED/table?dl=481DC, по состоянию на 5 февраля 2021 г.).

3.5 Линейчатая диаграмма

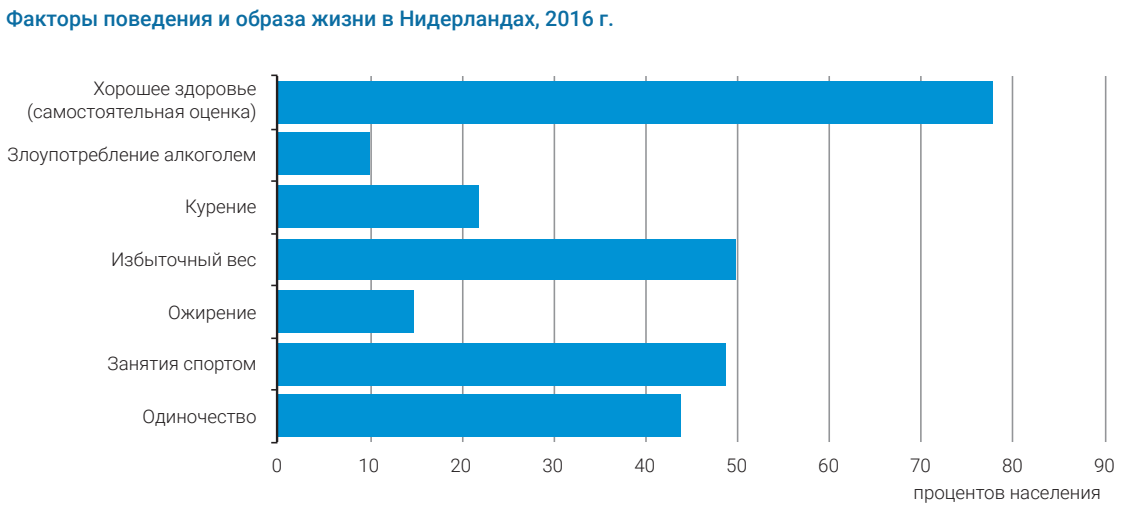
Линейчатая диаграмма лучше всего подходит для отображения распределения значений номинальных переменных. Обратите внимание, что на оси Y линейчатой диаграммы нельзя пропускать ни одну подпись. Это особенность номинальных переменных. Когда на столбчатой диаграмме отображаются порядковые переменные, как в случае рис. 5, подписи на оси X можно пропускать, хотя это и не идеальный вариант оформления.

Поскольку подписи, обозначающие номинальные переменные, могут быть длинными, линейчатая диаграмма в данном случае удобнее, чем столбчатая. Для удобства чтения подписи на оси Y выравнивают по правому краю.

Если для отображения номинальных переменных используется столбчатая диаграмма, подписи, идущие вдоль оси X, могут стать слишком длинными для размещения по горизонтали, и многие инструменты автоматически ставят их под углом. Текст, набранный по диагонали, выглядит неаккуратно и менее читаем, поэтому для номинальных данных более оптимальным решением является линейчатая диаграмма.

С технической точки зрения на линейчатой диаграмме оси X и Y расположены не так, как на столбчатой: ось Y ориентирована горизонтально, а ось X – вертикально.

Рис. 6. Пример линейчатой диаграммы

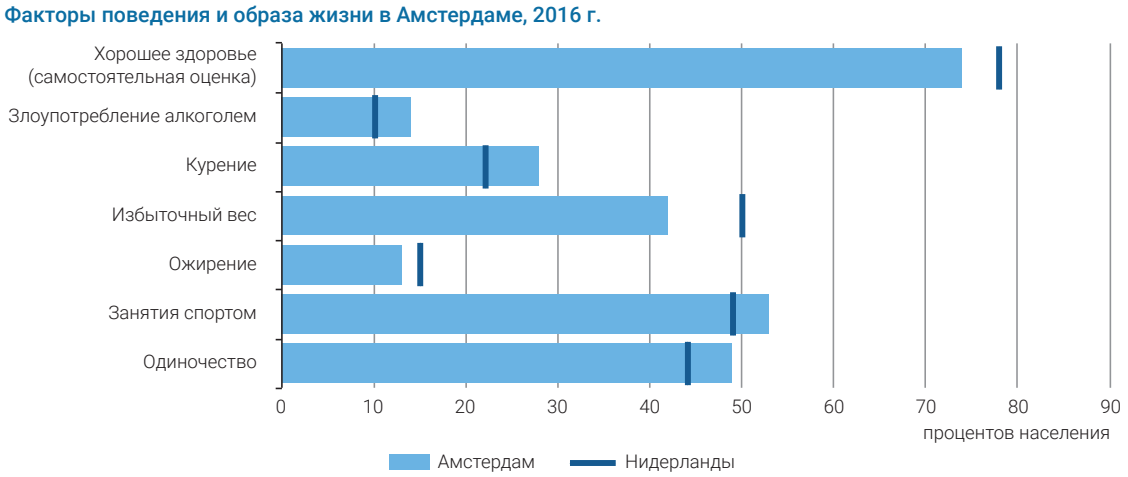


Источник данных: Gezondheid per wijk en buurt 2016 [онлайновая база данных] [на нидерландском языке]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2019 (https://statline.rivm.nl/#/RIVM/nl/dataset/50052NED/table?dl=481DE по состоянию на 5 февраля 2021 г.).

3.6 Диаграмма-шкала с маркером

Диаграмма-шкала с маркером – это линейчатая диаграмма с дополнительным **маркером** (засечкой) на каждой полосе. Такой маркер, например, может отмечать значение аналогичного показателя для другой группы населения или целевое значение для сравнения.

Рис. 7. Пример диаграммы-шкалы с маркером



Источник данных: Gezondheid per wijk en buurt 2016 [онлайновая база данных] [на нидерландском языке]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2019 (https://statline.rivm.nl/#/RIVM/nl/dataset/50052NED/table?dl=481DE по состоянию на 5 февраля 2021 г.).

3.7 Гистограмма

Гистограмма выглядит как линейчатая диаграмма, но отражает распределение частот, а не тренд на порядковой шкале. По оси X гистограммы перечислены **разряды** (интервалы) переменной; по оси Y отсчитывается частота, поэтому каждая полоса пропорциональна частоте соответствующего разряда.

Гистограмма – это приблизительное представление распределения числовых данных. Впервые она была предложена Карлом Пирсоном *(5)*. Для того чтобы построить гистограмму, необходимо сначала разделить диапазон значений на разряды – то есть выделить в нем серию интервалов, – а затем подсчитать, сколько значений попадает в каждый интервал. Разряды обычно определяются как последовательные непересекающиеся интервалы значений переменной. Они должны быть смежными и часто (но не обязательно) имеют одинаковый размер *(6)*.

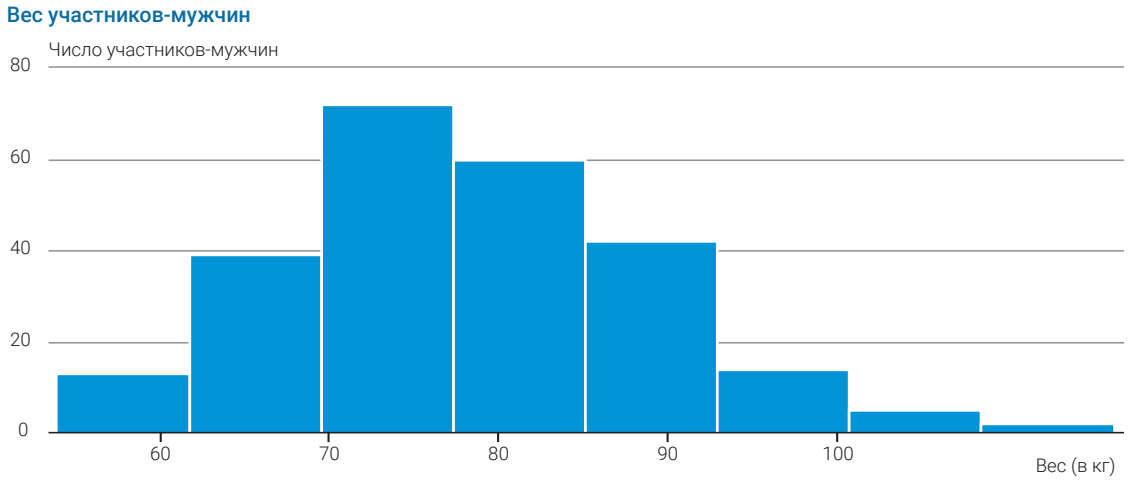
Поскольку гистограмма показывает частоту и, следовательно, визуализирует **распределение** переменной, столбцы обычно располагаются вплотную друг к другу без зазоров.

Разряды могут выделяться разной ширины и в разном количестве. Не существует «идеального» числа – разряды разного размера могут помочь выявить разные характеристики данных. Использование более широких интервалов при низкой плотности данных позволяет снизить шум, обусловленный случайностью выборки; использование более узких интервалов при высокой плотности (когда сигнал заглушает шум) позволяет оценить плотность более точно. Таким образом, в некоторых случаях имеет смысл сделать ширину разряда на гистограмме переменной. Однако активно используются и разряды равной ширины.

Некоторые теоретики пытались определить оптимальное количество разрядов, но эти методы обычно связаны со смелыми предположениями относительно формы распределения. В зависимости от фактического распределения данных и целей анализа может потребоваться разная ширина разряда, поэтому для определения подходящей ширины обычно требуются эксперименты. Однако существуют различные ценные соображения и практические правила.

Наиболее распространенный способ выбора количества разрядов – это **метод квадратного корня**. Согласно этому методу предлагается извлечь квадратный корень из числа элементов данных в выборке и округлить полученное значение до следующего целого числа. Этот метод используется для построения гистограмм в Excel и многих других программах. Изучение других методов при желании можно начать с материалов «Википедии» *(6)*.

Рис. 8. Пример гистограммы



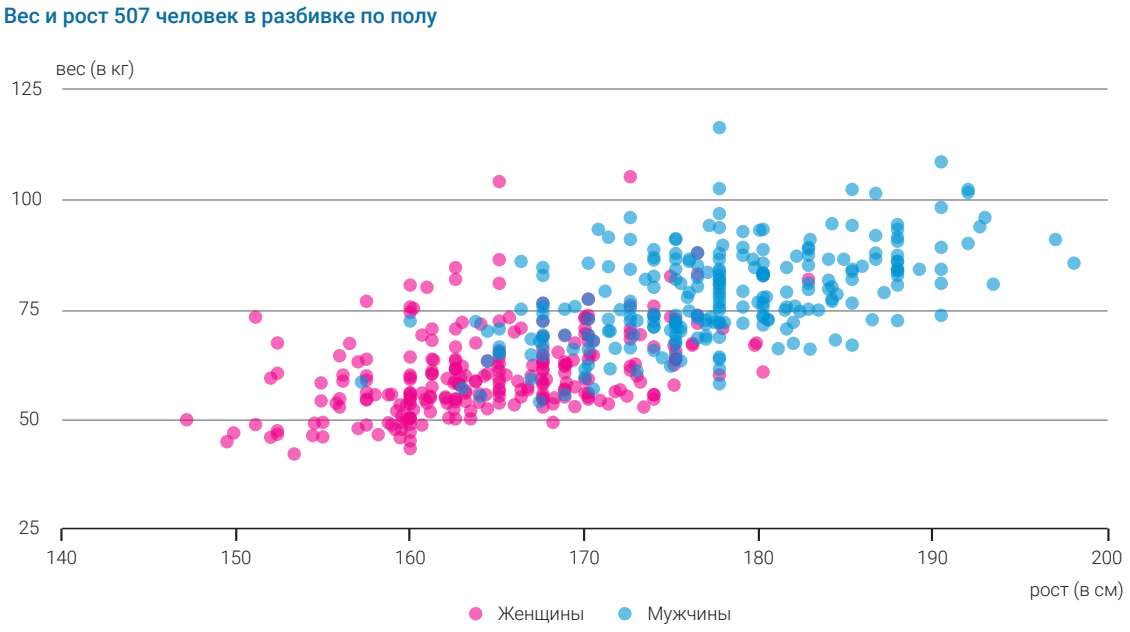
Источник данных: синтетический набор данных.

3.8 Точечная диаграмма

Точечные диаграммы (диаграммы рассеяния) используются для поиска корреляций. Каждая точка на точечной диаграмме означает «когда x равно этому, y равно тому». Таким образом, если в распределении точек наблюдается определенный тренд (вверх влево, вниз вправо и т. д.), между ними существует связь. Если точки полностью рассеяны и трендов не наблюдается, можно заключить, что переменные вообще не влияют друг на друга.

Точечная диаграмма на рис. 9 показывает взаимосвязь между весом и ростом 507 человек с разбивкой по полу. Она взята из исследования, посвященного изучению взаимосвязей между параметрами тела *(7)*.

Рис. 9. Пример точечной диаграммы

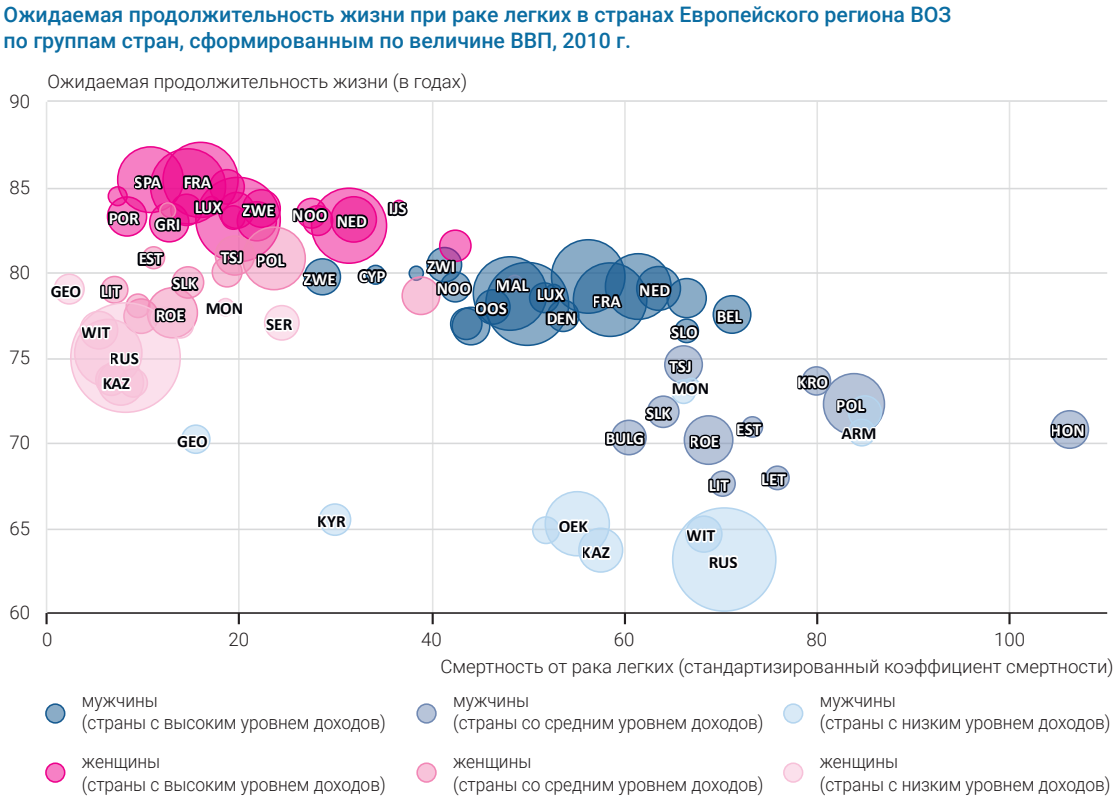


Источник данных: Heinz et al. *(7)*.

3.9 Пузырьковая диаграмма

Пузырьковая диаграмма – это вариант точечной диаграммы, где каждая точка изображена в виде «пузырька», площадь которого также несет определенную информацию в дополнение к положению точки по координатным осям. Трудность, связанная с пузырьковыми диаграммами, заключается в том, что пузырьки могут не помещаться на осях; поэтому не все данные подходят для этого типа визуализации.

Рис. 10. Пример пузырьковой диаграммы



Источник данных: Европейский портал информации здравоохранения: Путеводитель по базе данных «Здоровье для всех» [онлайновая база данных]. Копенгаген: Европейское региональное бюро ВОЗ; 2020 (https://gateway.euro.who.int/ru/hfa-explorer/, по состоянию на 5 февраля 2020 г.).

3.10 Круговая диаграмма

Круговая диаграмма – это способ иллюстрации процентных долей, поскольку она показывает каждый элемент как часть целого. Ее основное преимущество заключается в том, что идея об «отношении части и целого» отражена в ней самым непосредственным образом *(8)*.

Однако, несмотря на очевидность содержимого круговой диаграммы, линейчатые диаграммы гораздо лучше приспособлены для сравнения величин каждой из частей. Круговые диаграммы позволяют легко оценить величину сегмента, только когда она близка к 0%, 25%, 50%, 75% или 100%. Размер любых других долей на круговой диаграмме оценить трудно, однако их легко различить на линейчатой диаграмме – благодаря наличию шкалы *(8)*.

Линейчатую диаграмму в целом рекомендуется использовать вместо круговой. На рис. 11 показан один и тот же набор данных, визуализированный обоими способами. В этом примере на круговой диаграмме трудно оценить процент людей, умирающих от дыхательной недостаточности. Линейчатая диаграмма значительно упрощает эту задачу, сразу показывая, что дыхательной недостаточностью обусловлено около 7,5% общего числа случаев смерти.

На круговой диаграмме необходимо использовать разные цвета для разных сегментов, а это означает, что для визуализации данных может потребоваться множество цветов. На линейчатой диаграмме разные элементы могут иметь одинаковые цвета, хотя при необходимости можно выделить один элемент, назначив его полосе другой цвет, как в гистограмме на рис. 11.

Дополнительные доводы в пользу того, что круговая диаграмма может быть не лучшим способом визуализации, см. в эссе Стивена Фью *Save the pies for dessert [«Приберегите пироги на десерт»] (8)*.

Рис. 11. Пример круговой диаграммы и линейчатой диаграммы в качестве альтернативы

Источник данных: Levensverwachting. In: VTV-2018 [веб-сайт] [на нидерландском языке]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2018 (https://www.vtv2018.nl/Levensverwachting#bv1_3_1, по состоянию на 5 февраля 2021 г.).

3.11 Диаграмма наклона

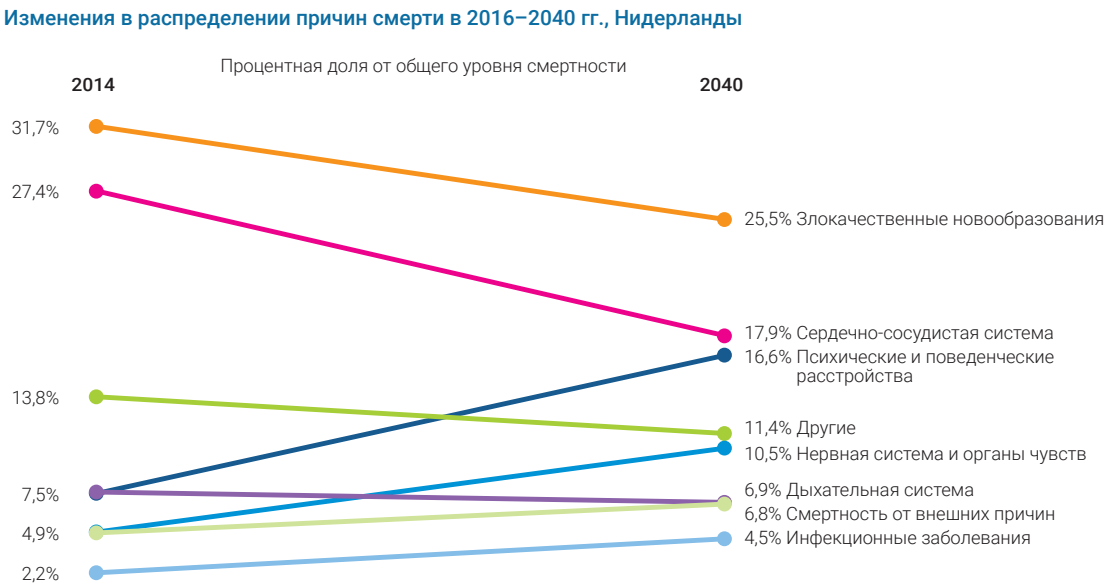
Диаграмма наклона – это брат-близнец линейного графика. На линейном графике отображаются три или более точки времени, а на диаграмме наклона – ровно две.

Диаграммы наклона – действительно эффективный инструмент для визуализации изменений со временем в наборе объектов (таких как страны, причины смерти или источники энергии). Они не только показывают изменение значений для одной сущности, но и сравнивают ее с другими сущностями в данном наборе.

Графики наклона определены Эдвардом Тафти в его книге 1983 *г. The visual display of quantitative information [«Визуальное отображение количественной информации»] (9).* Этот тип диаграммы полезен для демонстрации следующих аспектов:

- иерархические отношения набора сущностей в два момента времени (так, на рис. 12 сравнивается рейтинг причин смерти в 2014 и 2040 гг. (прогноз));
- конкретная процентная доля, связанная с каждой из сущностей в каждый из этих годов (см. проценты рядом с причинами смерти);
- изменение доли каждой из сущностей с течением времени (наклон линии каждой из причин смерти);
- темпы изменения доли каждой из сущностей в сравнении с тепами изменения других сущностей (сравнение наклонов между собой);
- любые заметные отклонения в общем тренде (обратите внимание на изменение линии «Психические и поведенческие расстройства» по сравнению с другими причинами смерти): отклоняющиеся от нормы наклоны.

Рис. 12. Пример диаграммы наклона



Источник данных: Levensverwachting. In: VTV-2018 [веб-сайт] [на нидерландском языке]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2018 (https://www.vtv2018.nl/Levensverwachting#bv1_3_1, по состоянию на 5 февраля 2021 г.).

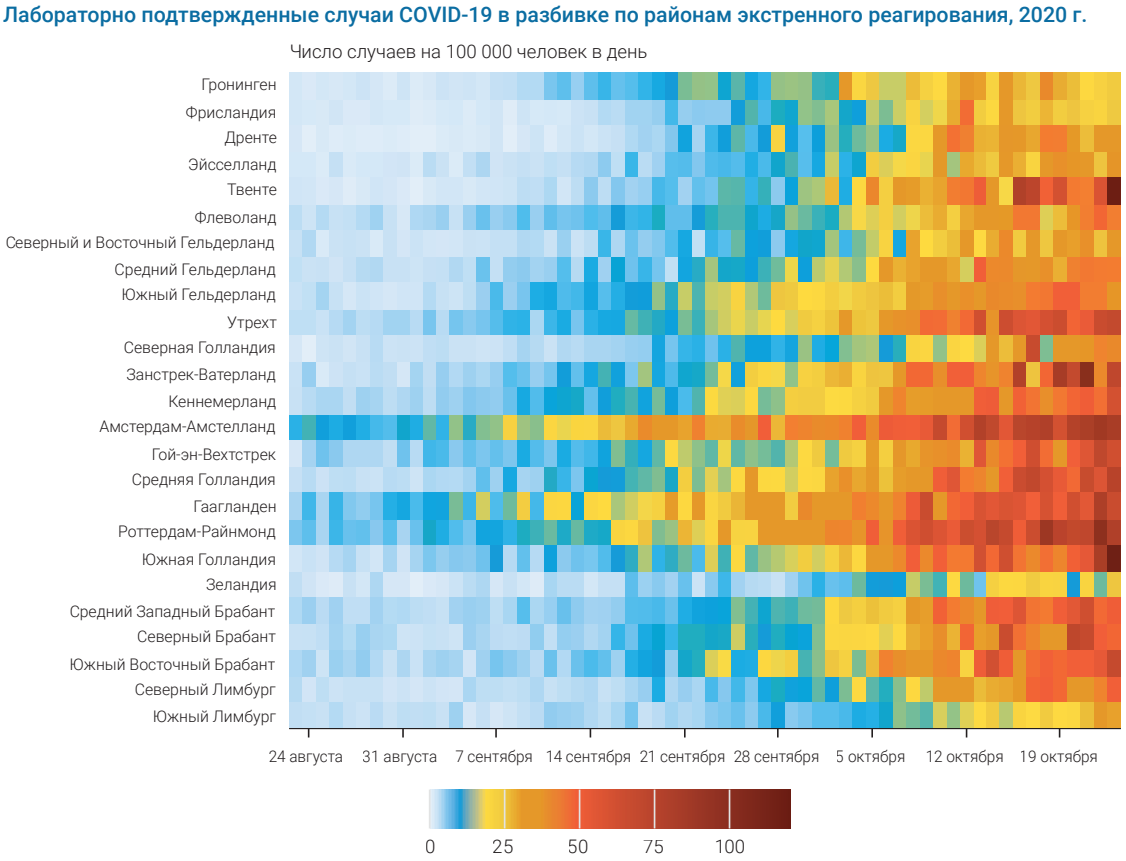
3.12 Тепловая карта

Тепловая карта по сути представляет собой матрицу с цветовым кодированием. Из названия можно предположить, что это карта, аналогичная тем, которые будут обсуждаться в разделе 3.13, но это не так. В действительности это матрица, где каждая ячейка закрашивается по определенной формуле. Оттенок цвета означает относительное значение или степень риска, соответствующий этой ячейке. Обычно используется расходящаяся цветовая схема, как в примере на рис. 13.

Этот тип визуализации полезен, потому что цвета легче интерпретировать, чем числа. Тепловая карта позволяет визуализировать множество хронологических последовательностей так, чтобы линии не мешали друг другу и не превращались в спагетти.

На рис. 13 показаны данные по 25 регионам. Эти данные также можно отобразить с помощью интерактивной **карты движения**, как описано в разделе 3.13. В таком случае будет подчеркиваться пространственное распределение, а изменение во времени станет второстепенным элементом. На тепловой же карте наибольшее внимание уделяется временны́м тенденциям, а пространственное измерение отходит на второй план.

Рис. 13. Пример тепловой карты

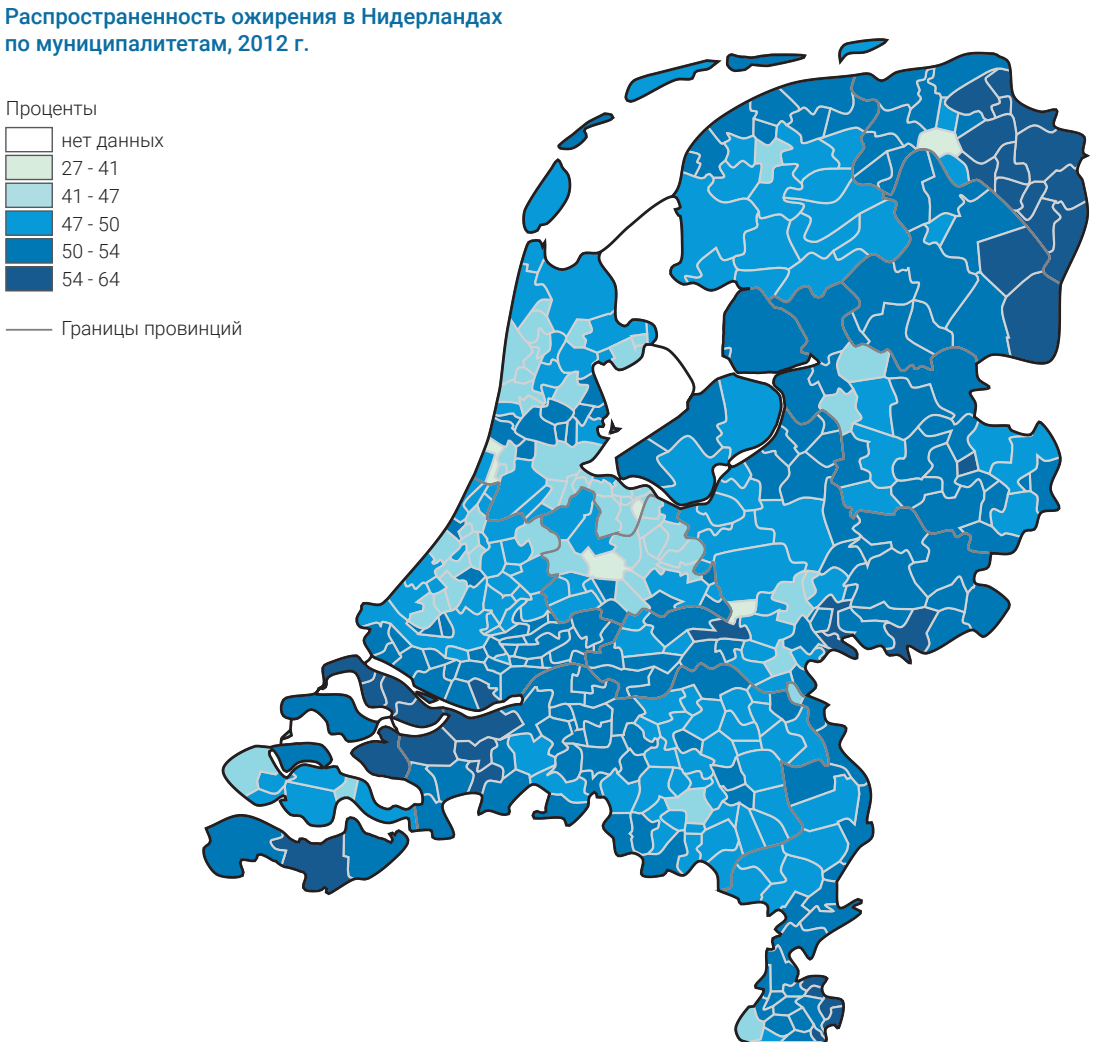


Источник данных: Covid-19 aantallen per gemeente per publicatiedatum [онлайновая база данных] [на нидерландском языке]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2020 (https://data.rivm.nl/geonetwork/srv/dut/catalog.search#/metadata/5f6bc429-1596-490e-8618-1ed8fd768427, по состоянию на 5 февраля 2021 г.).

3.13 Карта

Данные, с которыми работает эпидемиология, чаще всего имеют географический аспект, который упрощает их отображение на карте. На рис. 14 – пример карты, показывающей процент людей с избыточным весом в каждом муниципалитете Нидерландов.

Рис. 14. Пример карты-хороплета



Источник данных: Gezondheid per buurt, wijk en gemeente [онлайновая база данных] [на нидерландском языке]. Bilthoven: National Institute of Public Health and the Environment (RIVM); 2020 (https://www.rivm.nl/media/smap/index.html, по состоянию на 5 февраля 2021 г.).

Можно выделить два основных типа карт:

- топографические карты, используемые для навигации;
- тематические карты, используемые для визуализации пространственных закономерностей и процессов.

Тематическое картографирование – основной инструмент визуализации статистических данных. Среди множества различных видов тематических карт широко используются следующие.

- На **хороплетах** представляются статистические данные, агрегированные по заранее определенным регионам, таким как страны или штаты, путем окрашивания или затенения этих регионов (см. пример на рис. 14).
- На **хорохроматических картах** представляется пространственное распределение категориальной или номинальной переменной. Типичные примеры – геологические карты, карты почвы, растительности, землепользования, градостроительного зонирования и типов климата.
- **Карты изолиний** отображают непрерывные количественные поля, такие как карты осадков или высот, путем разделения пространства на области, каждая из которых содержит определенный диапазон значений поля. Таким образом, граница каждой области – изолиния – представляет собой совокупность координат, в которых переменная имеет какую-либо постоянную величину.
- В **картах с пропорциональными символами** используются точечные символы разных размеров (высота, длина, площадь или объем) для представления количественных статистических значений, связанных с отображенными на карте различными областями или местоположениями.
- На **картах плотности точек** по заданной территории размещаются маленькие точечные символы, обозначающие пространственное распределение рассматриваемого явления. Местоположение каждой точки может соответствовать фактическому местоположению одной единицы/случая. Примером карты плотности точек является знаменитая карта Джона Сноу (рис. 15; *(10)*).
- На **картах потоков** используются линейные символы для изображения движения или взаимосвязи между двумя или более географическими областями, например авиаперелетов, финансовой помощи или торговли.
- **Карты движения** интерактивны и могут использоваться только в интерактивной среде, такой как веб-страница или презентация PowerPoint. Изменения во времени показываются на такой карте как в кино. Иногда создается несколько небольших карт, каждая из которых относится к своему периоду времени, и при отображении на одной странице они сменяют друг друга, функционируя как карта движения.

Рис. 15. Карта вспышки холеры в округе Св. Джеймс, Вестминстер, осень 1854 г.

Источник: Snow *(10)*.

Дисциплина, которая занимается нанесением данных на карту, называется картографией. Стандартный справочник по картографии – Kraak & Ormeling *(11)*. Картография сама по себе является полноценной исследовательской областью, и обсуждение всех тонкостей составления и использования карт в целом и тематических карт в частности не входит в задачи данного документа. Для изучения тематического картографирования хорошей отправной точкой является соответствующая статья в «Википедии» *(12)*.

4. Важность пояснительных надписей

Пояснительные надписи нужны на каждой диаграмме: без них сложно понять представленные данные. Как отмечалось во введении, пояснительные надписи не так важны на этапе исследования, когда конечный пользователь диаграмм – сам аналитик или его коллеги, с которыми он общается напрямую.

Однако, когда речь идет о широкой публикации материалов, пояснительные надписи чрезвычайно важны. Во время подготовки визуализации, использовавшейся на этапе исследования, для презентации публике, важно также добавлять и улучшать пояснительные надписи. Следует рассмотреть ряд элементов:

- заголовок и (факультативно) подзаголовок диаграммы;
- название оси X;
- название оси Y;
- элементы легенды;
- ссылка на источник;
- авторство;
- пояснения.

4.1 Заголовок и подзаголовок

Заголовок диаграммы – это первое, что видит пользователь, обращаясь к визуализации данных, поэтому это крайне важный элемент. Визуализация должна быть простой для понимания и не должна опираться лишь на числа: суть чисел должна быть изложена словами. Без соответствующего заголовка маркеры, линии и числа могут означать что угодно.

Заголовок дает информацию о предмете диаграммы. Если применимо, в нем также должны указываться область (например, страна, штат, провинция) и период (например, год, месяц, временной диапазон).

Заголовки должны быть четкими и краткими, без лишних слов. Например, вот два заголовка:

- Среднее количество госпитализаций в сутки среди жителей Бухары, Узбекистан, в октябре 2019 г.
- Количество госпитализаций в сутки в Бухаре, Узбекистан, октябрь 2019 г.

Второй вариант передает требуемую информацию четко и без лишних слов.

Если заголовок содержит всю необходимую информацию, подзаголовок не требуется. Подзаголовок можно использовать, если заголовок оказался слишком длинным, но обычно можно просто укоротить заголовок. Возможно также выносить в подзаголовок указание географической области и/или периода. Если делать это последовательно, пользователю будет проще найти нужную информацию.

4.2 Название оси X

Название оси X следует указывать всегда, если только назначение оси X не совсем очевидно. В заголовке оси X должен указываться параметр и, если применимо, единицы измерения, в которых он измеряется, например:

- масса (кг);
- длина (км);
- возрастные категории.

Часто на оси X отмечается период времени, исчисляемый в годах или месяцах. В таких случаях очевидно, что речь идет о временном периоде, и заголовок оси X можно опустить. В случае номинальной шкалы заголовок оси X не требуется в принципе, поскольку он будет дублировать сведения из заголовка диаграммы или элементов легенды.

4.3 Название оси Y

Название оси Y должно присутствовать на диаграмме всегда. Оно поясняет, в каких единицах измеряются представленные на диаграмме данные, например:

- число;
- число (тыс.);
- проценты;
- евро (млн);
- евро на жителя;
- годы.

Существует соблазн повторить в заголовке оси Y тему диаграммы. Если в этом нет крайней необходимости, следует избегать дублирования информации в диаграмме и в пояснительных надписях.

4.4 Элементы легенды

Если диаграмма содержит только один ряд данных, в легенде нет необходимости, так как будет очевидно, какой ряд данных представлен. Однако если диаграмма содержит два ряда данных или более, легенда необходима. Каждый ряд данных должен быть представлен соответствующим элементом легенды.

Рекомендуется формулировать каждый элемент легенды коротко и понятно, чтобы не загромождать иллюстрацию.

4.5 Ссылка на источник

Сведения об источнике информации чрезвычайно важны для проверки качества и надежности представленных данных. Например, если данные поступают от национального статистического бюро, ВОЗ, Организации экономического сотрудничества и развития или другой надежной организации, диаграммы вызывают большее доверие, чем если данные получены из некого малоизвестного источника в Сети. В последнем случае вполне вероятно, что результаты будут оспариваться.

4.6 Авторство

Иногда в пояснительных надписях полезно также упомянуть организацию, которая разработала диаграмму. Если диаграмма используется вне оригинального контекста, важно, чтобы она содержала информацию о ее авторстве. Особенно часто диаграммы используются в отрыве от исходного материала в интернете.

4.7 Пояснения

Наконец, всегда есть возможность добавить дополнительный текст в области построения диаграммы. Например, если в тренде наблюдается разрыв ввиду того, что в разные годы использовались разные методы измерения, стоит рассмотреть возможность пояснить это текстом на диаграмме.

5. Передовая практика создания диаграмм

5.1 Прямое подкрепление тезисов

Первый совет – наиболее важный: диаграммы не должны оставлять «скрытых картинок», требующих от читателя экстраполировать информацию; визуальные материалы должны прямо соответствовать ключевым тезисам текста. Может показаться, что это очевидно, но иногда возникает соблазн нарушить это правило. Поясним это на следующем простом примере.

На рис. 16 показан один из способов визуализации разницы в продолжительности жизни мужчин и женщин в Нидерландах на протяжении определенного периода. Этот график подтверждает наблюдение, что с 1950 г. разница увеличивалась, а с 2000 г. уменьшалась.

Однако продолжительность жизни мужчин и женщин показывается на нем независимо. Увеличение и уменьшение разницы доступны нам лишь косвенно, в форме некого визуального вывода. Если наша цель – подчеркнуть разницу, лучше визуализировать саму разницу, как на рис. 17. На этом рисунке точно отражены максимальные и минимальные различия в продолжительности жизни между мужчинами и женщинами. Так гораздо проще увидеть, когда рост разницы прекратился и началось ее снижение.

Конечно, рис. 16 – вполне привлекательная диаграмма, содержащая полезную информацию для рассказа о различиях в тенденциях роста между мужчинами и женщинами. Она была бы полезна, если бы нам необходимо было проиллюстрировать рост продолжительности жизни в Нидерландах начиная с 1950 г. Пригодилась бы она и в качестве основы для отчета о том, как замедлялся рост продолжительности жизни мужчин и женщин в разные периоды. Однако изменение разницы в продолжительности жизни мужчин и женщин с течением времени лучше всего покажет рис. 17.

Рис. 16. Пример диаграммы со «скрытой картинкой»

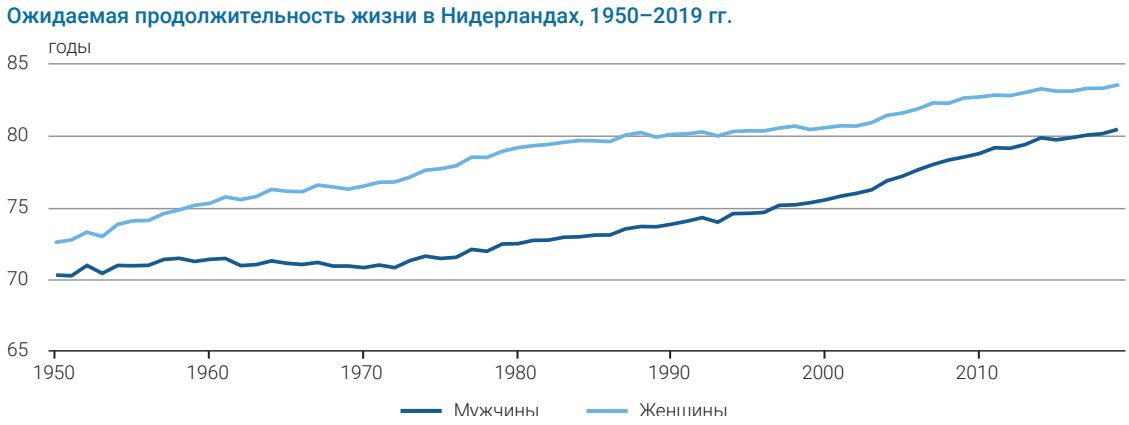
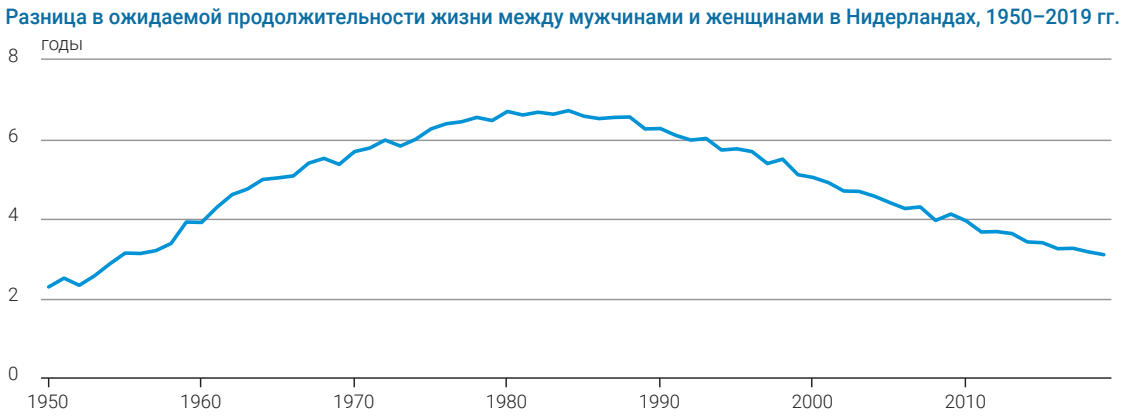


Рис. 17. Пример диаграммы, прямо подкрепляющей ключевой тезис



Источник данных: Levensverwachting; geslacht, leeftijd (per jaar en periode van vijf jaren) [онлайновая база данных] [на нидерландском языке]. The Hague: Statistics Netherlands (CBS); 2020 (https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37360ned/table?dl=481ED, по состоянию на 5 февраля 2020 г.).

5.2 Избегайте графмусора

Обратите внимание, что легенда на рис. 17 не требуется. Это подводит нас ко второму совету: избегайте **«графмусора»**. Так мы называем все визуальные элементы в диаграммах, которые не требуются для понимания представленной информации или отвлекают зрителя от нее *(13)*.

Термин графмусор (chartjunk) был введен Эдвардом Тафти в его книге 1983 г. *The visual display of quantitative information [«Визуальное отображение количественной информации»] (9)*:

> *При оформлении графиков возникает много декоративных элементов, которые не говорят зрителю ничего нового. Цель такого украшательства бывает разной: сделать так, чтобы график выглядел более научным и точным, «оживить» картинку, дать дизайнеру возможность проявить свои художественные способности. Но, независимо от причины, это все чернила без данных, или же избыточные чернила для данных, и часто это просто графмусор.*

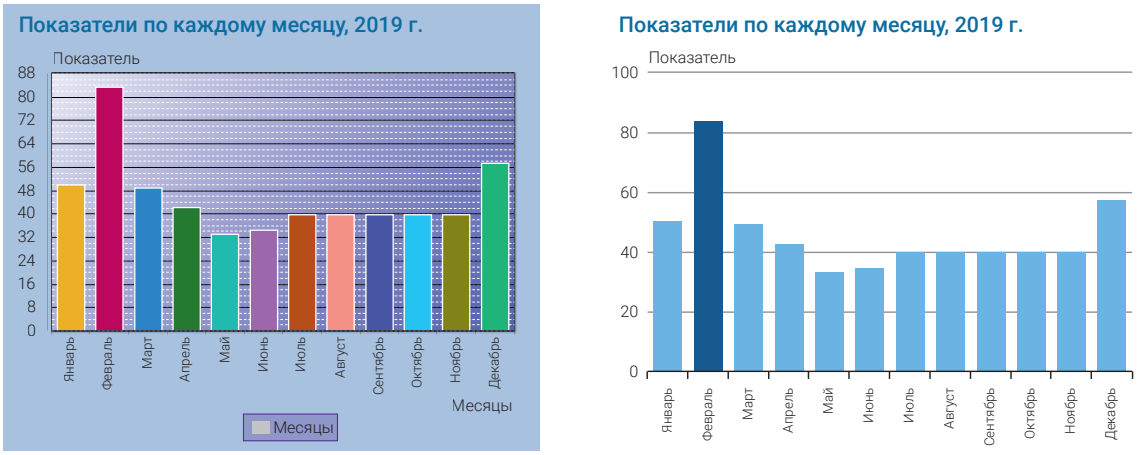
Графмусор тесно связан с **соотношением данные / чернила –** еще одним термином, введенным Тафти. Чернила для данных – это нестираемые чернила, используемые для представления данных. Если удалить из иллюстрации чернила для данных, она потеряет свое содержание. В свою очередь, чернила без данных – это чернила, не передающие никакой информации.

Соотношение данные / чернила – это доля чернил (или пикселей, если речь идет об информации на экране), реально используемых для представления данных, без какой-либо избыточности, по отношению к общему количеству чернил (или пикселей), используемых во всей иллюстрации, например в таблице или на графике. Наша цель – разработать иллюстрацию с максимально возможным соотношением объема используемых чернил к данным (то есть максимально приближенным к 1,0 или 100%), не исключив ничего нужного для эффективного донесения необходимой информации.

Левая диаграмма на рис. 18 – это пример очень низкого соотношения данные / чернила. Много чернил потрачено на крупный шрифт, на фон диаграммы и фон легенды, на саму легенду, на уникальные цвета каждого элемента данных и на границу столбцов. Подпись оси X также не нужна, а линии сетки слишком толстые и слишком плотные. Можно сэкономить много чернил.

Правая же диаграмма на рис. 18 – это пример чистого оформления диаграммы с одним выделенным столбцом. Ее можно использовать как наглядную иллюстрацию для материала, в котором говорится о достижении максимального результата в феврале. В этом случае диаграмма поддерживает ключевой тезис и использование дополнительных «чернил для данных» вполне оправданно.

Рис. 18. Пример диаграммы, наполненной графмусором, и чистой диаграммы



Источник данных: синтетический набор данных.

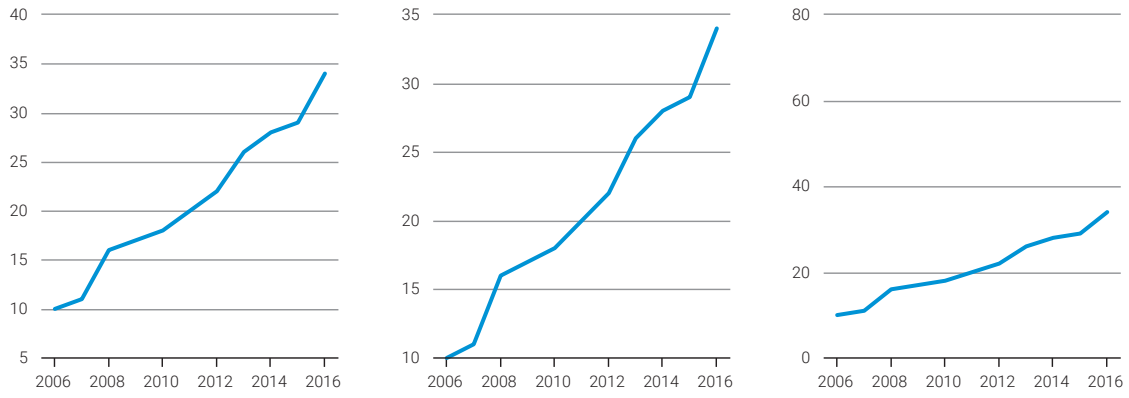
5.3 Точка отсчета оси Y и соотношение сторон диаграммы

Обратите внимание, что ось Y линейного графика не обязательно должна начинаться с нуля (см. рис. 16). В случае ожидаемой продолжительности жизни интерес представляют

даже незначительные изменения с течением времени. Если бы ось Y начиналась с нуля, нужная информация не была бы доведена до аудитории. С другой стороны, если ось Y будет слишком сжата, изменения могут выглядеть преувеличенными. В 1954 г. Даррелл Хафф опубликовал свою знаменитую книгу «Как лгать при помощи статистики» *(14)*. В этом небольшом издании приведены многочисленные примеры, показывающие, что происходит, если к построению диаграмм подходить неправильно. В настоящем разделе обсуждаются некоторые основные проблемы, которых можно легко избежать.

Необдуманное изменение масштаба оси Y может привести к различным недоразумениям. Растянув или сжав ее, мы преувеличим или преуменьшим значимость изменений *(15)*. Смещая минимальную и максимальную точку диаграммы, мы также влияем на то, как выглядит график. Растянув график по высоте, мы можем придать картине необоснованный драматизм, а увеличив его по ширине – сделать некоторые изменения недостаточно яркими. См. пример на рис. 19.

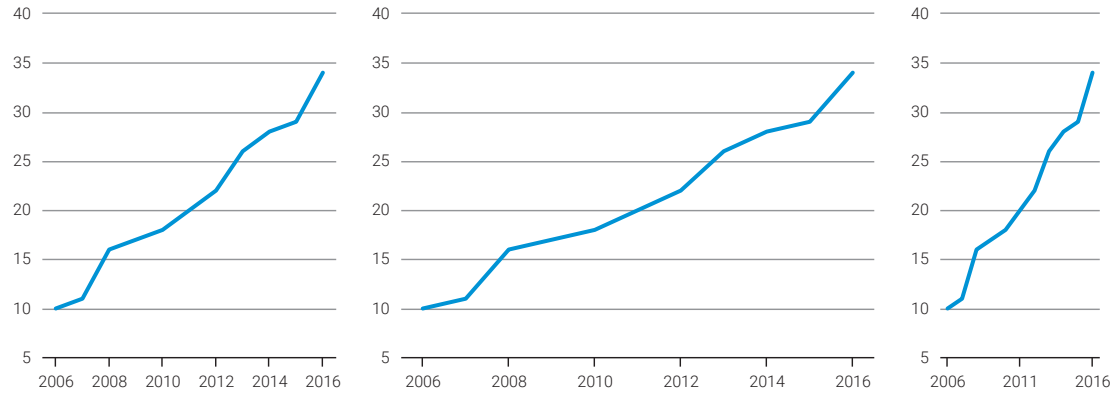
Рис. 19. Примеры разных настроек оси Y при тех же данных



Источник данных: синтетический набор данных.

Как и изменения, показанные на рис. 19, изменение соотношения сторон диаграммы тоже влияет на то, как график выглядит. На рис. 20 показано влияние различных соотношений ширины/высоты диаграммы на ее вид.

Рис. 20. Примеры изменения соотношения сторон диаграммы



Источник данных: синтетический набор данных.

Как и в случае с растянутой или сжатой фотографией, размер диаграммы или соотношение ее сторон могут повлиять на вид представленной информации. Если неправильное соотношение сторон на фотографии обычно сразу бросается в глаза, искажение диаграммы легко может остаться незамеченным. Независимо от того, приводит ли это к преувеличению или преуменьшению представленных данных, это вводит аудиторию в заблуждение.

«Нет единого правила в отношении того, насколько высоким или широким должен быть график, но полезно взять за ориентир угол в **45°** – средний угол наклона на вашей диаграмме должен приближаться к 45°», – говорит специалист по визуализации данных Энди Кирк *(16)*. Пытаться точно измерить этот угол может быть затруднительно, но оценить на глаз – вполне возможно. Хорошее обоснование правила 45° можно найти в работе Уильяма Кливленда *(17)*.

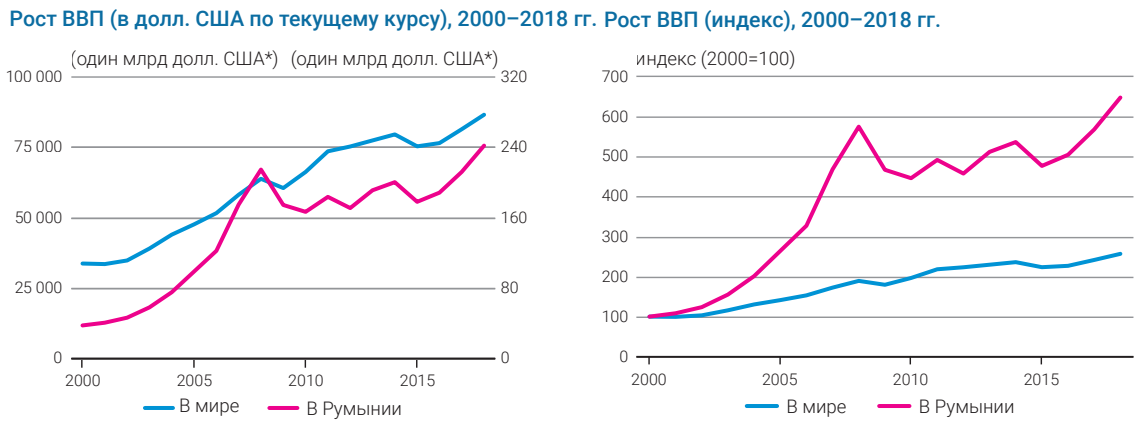
5.4 Избегайте использования двух осей Y на одном графике

Относительно просто и иногда полезно визуализировать несколько наборов данных с общей горизонтальной осью, если все данные выражены в одних и тех же единицах измерения. Однако, если используются разные единицы измерения, необходимо добавить дополнительную вертикальную ось справа от диаграммы. Отображение двух наборов данных с одной шкалой слева и другой справа может сбить с толку; это также может создать видимость несуществующих отношений.

Для того чтобы расшифровать такую диаграмму и понять, какие данные следует читать относительно какой оси, требуются время и усилия. Даже если читатель решит эту загадку, он может попытаться сравнить величины значений между двумя наборами данных, но это бессмысленно, учитывая, что шкалы и единицы измерения разные.

Как правило, нежелательно втискивать слишком много информации в небольшое пространство. Для того чтобы рассказать сложную историю, лучше использовать две диаграммы или более. Если данные необходимо разместить на одной диаграмме, можно использовать индексацию. Исчерпывающий обзор проблем, которые возникают при использовании двух осей Y, и их решений можно найти в блоге Лизы Рост *(18)*.

Рис. 21. Пример отображения двух наборов данных на одной диаграмме



Источник данных: источник: Всемирный банк

5.5 Избегайте трехмерных диаграмм

Такие инструменты, как Microsoft Excel, позволяют легко создавать трехмерные диаграммы, поэтому они популярны. Проблема в том, что проекция трехмерных объектов на плоскость для печати или отображения на мониторе неизбежно приводит к искажению данных. Человеческий зрительный аппарат пытается исправить это искажение, переводя двухмерную проекцию 3D-изображения обратно в трехмерное пространство, но эта коррекция лишь частична.

Восприятие данных на трехмерной диаграмме всегда искажено, поэтому их лучше избегать. В главе 26 книги Клауса Уилке *Fundamentals of data visualization [«Основы визуализации данных»] (19)* приводятся хорошие примеры искажения данных при визуализации с использованием 3D-диаграмм.

5.6 Избегайте круговых диаграмм

Как уже говорилось в разделе 3.10, круговая диаграмма редко бывает оптимальным вариантом. По круговым диаграммам лицам, принимающим решения, сложнее понять ключевые тезисы, заключенные в них. Эти диаграммы кажутся простыми и понятными, но на самом деле их трудно читать, и в большинстве случаев есть более удобная альтернатива.

Дети учат дроби, представляя, как пирог режут на равные части. Но когда сектора не равны – как это часто бывает с реальными данными – точно представить части целого, которые изображены на круговой диаграмме, бывает затруднительно. Человеческий мозг не умеет точно оценивать или сравнивать углы. Когда доли круга близки по размеру, сказать, какая из них больше, может быть трудно или вообще невозможно; когда же они заметно отличаются, можно легко определить, какая доля больше, но насколько именно больше сказать по-прежнему нельзя.

Лучший выбор для таких целей – линейчатая диаграмма. Люди от природы не очень хорошо умеют оценивать размер секторов круга, особенно с первого взгляда; они гораздо лучше приспособлены к тому, чтобы замечать различия в прямоугольных формах. Кроме того, практически невозможно сравнивать похожие сектора на двух разных круговых диаграммах. Для этого лучше всего подходят диаграммы наклона.

Последний аргумент в пользу отказа от круговых диаграмм касается дизайна: круговая диаграмма занимает гораздо больше места, чем другие варианты представления аналогичного набора данных. Кроме того, надписи не выстраиваются в одну линию, поэтому диаграмма может оказаться перенасыщенной и трудночитаемой. Извлечь точные данные из круговой диаграммы – нелегкий труд; например, приходится полагаться на подписи секторов, которые могут не помещаться в отведенном пространстве, или же обращаться к легенде, постоянно переводя взгляд между диаграммой и легендой.

6. Единство оформления диаграмм

Еще один важный момент – это единство оформления диаграмм. Именно наличие или отсутствие единого стиля может определить, будет отчет производить впечатление ясности или визуального шума. Визуальный шум может отвлечь внимание от сути материала, поэтому особенно важно использовать (или даже самостоятельно разработать) руководство по стилю визуализации данных *(20)*.

Единство стиля заключается в небольших, но важных деталях, в том числе в следующих.

- На каждой диаграмме должен использоваться одинаковый шрифт и цвет текста.
- Для каждого типа пояснительных надписей следует использовать соответствующий его важности размер шрифта. Заголовок диаграммы – наиболее важный элемент; за ним следуют подзаголовок, названия осей X и Y, подписи, размещенные вдоль обеих осей, и, наконец, ссылка на источник.
- Поля вокруг диаграмм и содержащихся в них элементов должны быть одинаковыми для всех диаграмм.
- Расположение заголовков и подписей должно быть одинаковым.
- Способ изображения осей X, Y, засечек на обеих осях и линий сетки должен быть по возможности унифицирован. Однотипные элементы на диаграммах должны иметь одинаковый цвет, ширину, длину и стиль.
- Использование цвета должно быть унифицировано. Важно отметить, что фирменные цвета организации обычно не подходят для визуализации данных *(21)*.
- Параметры элементов диаграмм должны быть унифицированы (толщина линий в линейных графиках, цвет и ширина границ столбцов и полос).

В иллюстрациях к настоящим рекомендациям все эти элементы унифицированы.

7. Дополнительная литература

В интернете можно найти множество дополнительных материалов по визуализации данных. Существует так много статей, блогов, книг и образцов, что решить с чего начать может стать непростой задачей. Поэтому в этом разделе мы приводим некоторые хорошие отправные точки для дальнейшего изучения темы.

Книги

Большинство книг доступны лишь на платной основе, но есть некоторые исключения.

- Полезная книга с хорошими примерами, которая отлично подходит в качестве бесплатного дополнительного чтения:
  Wilke CO. Fundamentals of data visualization: a primer on making informative and compelling figures. Sebastopol, CA: O'Reilly Media; 2019 (https://clauswilke.com/dataviz/).
- Чуть более старая (последнее обновление было в 2012 г.), но все же заслуживающая упоминания книга:
  Principles of epidemiology in public health practice. Atlanta, GA: Centers for Disease Control and Prevention; 2012 (https://www.cdc.gov/careerpaths/k12teacherroadmap/classroom/principlesofepi.html).

Четвертый урок в ней посвящен представлению данных общественного здравоохранения; в нем можно найти ответы на некоторые дополнительные вопросы.

Блоги

В интернете можно найти множество блогов, где можно почерпнуть хорошие идеи.

- Nightingale – журнал Общества визуализации данных. В нем публикуются полезные записи блогов и статьи о визуализации данных, некоторые из которых могут послужить очень ценным источником идей и вдохновения: https://medium.com/nightingale
- Chartable – блог разработчиков онлайн-инструмента для построения диаграмм и карт Datawrapper. Как и в Nightingale, здесь есть статьи о визуализации данных, в которых можно найти много интересного: https://blog.datawrapper.de/
- Perceptual Edge – официальный блог Стивена Фью, известного своими идеями о визуализации данных. С 2006 г. он опубликовал множество статей по широкому кругу вопросов, связанных с визуализацией данных: http://www.perceptualedge.com/

Библиография[2]

1. Few S. What is data visualization? Perceptual Edge [блог]. 4 May 2017 (https://www.perceptualedge.com/blog/?p=2636).[2]

2. What is data visualization and why is it important? [веб-сайт]. Campbell, CA: Import.io; 2019 (https://www.import.io/post/what-is-data-visualization/).

3. Anscombe FJ. Graphs in statistical analysis. Am Stat. 1973;27(1):17–21. doi:10.1080/00031305.1973.10478966.

4. Stevens SS. On the theory of scales of measurement. Science. 1946; 103(2684):677–80 (https://psychology.okstate.edu/faculty/jgrice/psyc3214/Stevens_FourScales_1946.pdf).

5. Pearson K. X. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. Philos Trans R Soc Lond A. 1895;186:343–414. (https://bayes.wustl.edu/Manual/Pearson_1895.pdf).

6. Histogram. In: Wikipedia, the free encyclopedia [веб-сайт]. Wikipedia; 2020 (https://en.wikipedia.org/wiki/Histogram).

7. Heinz G, Peterson LJ, Johnson RW, Kerk CJ. Exploring relationships in body dimensions. J Stat Educ. 2003;11(2). doi:10.1080/10691898.2003.11910711.

8. Few S. Save the pies for dessert. Perceptual Edge Newsletter. August 2007. (https://www.perceptualedge.com/articles/visual_business_intelligence/save_the_pies_for_dessert.pdf).

9. Tufte ER. The visual display of quantitative information. Cheshire, CT: Graphics Press; 1983.

10. Snow J. Map on the cholera outbreak in the Parish of St. James, Westminster, during the autumn of 1854. London; 1854 (https://commons.wikimedia.org/wiki/File:A_map_taken_from_a_report_by_Dr._John_Snow_Wellcome_L0072917.jpg; Creative Commons CC-BY 4.0 license).

11. Kraak MJ, Ormeling F. Cartography: visualization of geospatial data, fourth edition. Boca Raton, FL: CRC Press; 2020.

12. Thematic map. In: Wikipedia, the free encyclopedia [веб-сайт]. Wikipedia; 2020. (https://en.wikipedia.org/wiki/Thematic_map).

2 Все ссылки приведены по состоянию на 5 февраля 2021 г.

13. Chartjunk. In: Wikipedia, the free encyclopedia [веб-сайт]. Wikipedia; 2020 (https://en.wikipedia.org/wiki/Chartjunk).

14. Дарелл Хафф. Как лгать при помощи статистики. М.: Альпина Паблишер, 2015.

15. Misleading graph. In: Wikipedia, the free encyclopedia [веб-сайт]. Wikipedia; 2020 (https://en.wikipedia.org/wiki/Misleading_graph).

16. Kirk A. Data visualization: a handbook for data driven design. London: Sage; 2016.

17. Cleveland WS, McGill ME, McGill R. The shape parameter of a two-variable graph. J Am Stat Assoc. 1988;83(402):289–300. doi:10.2307/2288843.

18. Rost LC. Why not to use two axes, and what to use instead. Chartable [блог]. 8 May 2018 (https://blog.datawrapper.de/dualaxis/).

19. Wilke CO. Fundamentals of data visualization: a primer on making informative and compelling figures. Sebastopol, CA: O'Reilly Media; 2019 (https://clauswilke.com/dataviz/).

20. Cesal A. What are data visualization style guidelines? Nightingale [блог]. 10 July 2019 (https://medium.com/nightingale/style-guidelines-92ebe166addc).

21. Cesal A. How to create brand colors for data visualization style guidelines: your brand colors don't work for data visualization. Nightingale [блог]. 13 July 2020 (https://medium.com/nightingale/how-to-create-brand-colors-for-data-visualization-style-guidelines-dbd69c586dd9).

Всемирная организация здравоохранения

Европейское региональное бюро

Европейское региональное бюро ВОЗ

Всемирная организация здравоохранения (ВОЗ) – специализированное учреждение Организации Объединенных Наций, созданное в 1948 г., основная функция которого состоит в решении международных проблем здравоохранения и охраны здоровья населения. Европейское региональное бюро ВОЗ является одним из шести региональных бюро в различных частях земного шара, каждое из которых имеет свою собственную программу деятельности, направленную на решение конкретных проблем здравоохранения обслуживаемых ими стран.

Государства-члены

Австрия
Азербайджан
Албания
Андорра
Армения
Беларусь
Бельгия
Болгария
Босния и Герцеговина
Венгрия
Германия
Греция
Грузия
Дания
Израиль
Ирландия
Исландия
Испания
Италия
Казахстан
Кипр
Кыргызстан
Латвия
Литва
Люксембург
Мальта
Монако
Нидерланды
Норвегия
Польша
Португалия
Республика Молдова
Российская Федерация
Румыния
Сан-Марино
Северная Македония
Сербия
Словакия
Словения
Соединенное Королевство
Таджикистан
Туркменистан
Турция
Узбекистан
Украина
Финляндия
Франция
Хорватия
Черногория
Чехия
Швейцария
Швеция
Эстония

Всемирная организация здравоохранения
Европейское региональное бюро
UN City, Marmorvej 51
DK-2100 Copenhagen Ø, Denmark
Тел.: +45 45 33 70 00;
факс: +45 45 33 70 01
Эл. адрес: eurocontact@who.int
Веб-сайт: www.euro.who.int