Создание столбчатой диаграммы в R

Большинство пользователей (и я в том числе) применяют Excel для построения диаграмм. Однако, возможности Excel относительно ограничены, и в какой-то момент вам вполне вероятно захочется большего. Рассмотрим пример создания столбчатой диаграммы в статистическом пакете R и последующей обработки диаграммы в Adobe Illustrator для придания издательского качества вашей работе.[1] Если вы никогда не сталкивались с R, рекомендую начать с книги Алексея Шипунова Наглядная статистика. Используем R!

В качестве исходных данных возьмем результаты Нейтановского турнира по поеданию хот-догов за последние три десятилетия (рис. 1).

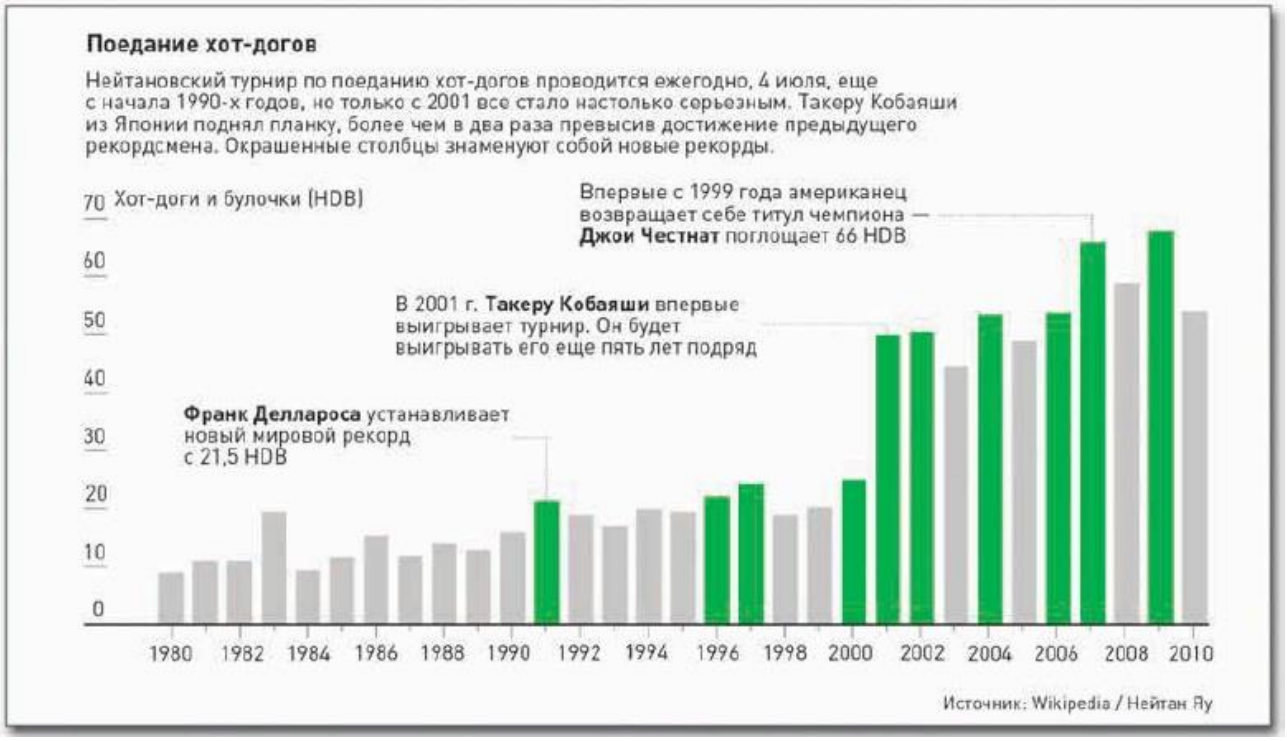


Рис. 1. Окончательный вариант диаграммы, которую вы будете стараться построить; чтобы увеличить картинку, кликните на ней правой кнопкой мыши и выберите опцию *Открыть картинку в новой вкладке*

В Википедии данные представлены в виде HTML-таблицы и включают в себя год, имя победителя, количество съеденных хот-догов, а также страну, которую представлял победитель. Данные были собраны в CSV- файл (рис. 2).

Рис. 2. Первые строчек данных в CSV- файле

[1] Заметка написана на основе материалов книги Нейтан Яу. Искусство визуализации в бизнесе. – М.: Манн, Иванов и Фербер, 2013. – С. 109–121.

Чтобы загрузить данные в R, используйте команду read.csv(). Вы можете загрузить файл с ПК:

```
# Change directory
setwd("D:\\Dropbox\\!Сайт\\7_Библиотека\\Яу\\Столбчатая диаграмма")

# Load data
hotdogs <- read.csv("data/hot-dog-contest-winners.csv", sep=",", header=TRUE)
```

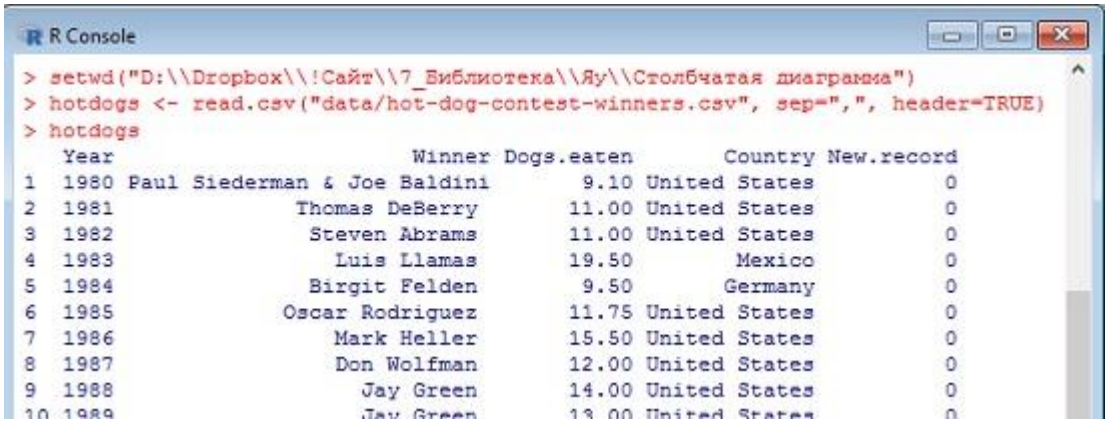
или обратиться к нему по URL:

```
hotdogs <- read.csv("http://datasets.flowingdata.com/hot-dog-contest-winners.csv", sep=",", header=TRUE)
```

Приложенный файл hot-dog-contest-winners.csv я дополнил данными за 2011–2015 гг. (на сайте же последний год – 2010).

Разберем фрагмент кода подробнее. Команда read.csv() загружает данные. У нее есть три аргумента. Первый — это местоположение данных. В приведенных примерах это папка "data" или URL. Второй аргумент, sep, определяет, какой знак разделяет колонки в файле с данными. В нашем случае это файл с разделителями-запятыми. Если бы это был файл, в котором разделитель – табуляции, тогда бы следовало указать \t. Последний аргумент — header — сообщает R, что у файла с данными есть шапка с названиями колонок. Последняя колонка – признак рекорда. Если в соответствующем году мировой рекорд оказался побит, в ячейке будет стоять 1, в противном случае — 0.

Итак, данные загружены в R и доступны через переменную hotdogs. Точнее говоря, данные сохранены в виде таблицы. Вот как она будет выглядеть, если вы наберете hotdogs (рис. 3):



```
R Console
> setwd("D:\\Dropbox\\!Сайт\\7_Библиотека\\Яу\\Столбчатая диаграмма")
> hotdogs <- read.csv("data/hot-dog-contest-winners.csv", sep=",", header=TRUE)
> hotdogs
   Year                      Winner Dogs.eaten       Country New.record
1  1980 Paul Siederman & Joe Baldini       9.10 United States          0
2  1981              Thomas DeBerry      11.00 United States          0
3  1982               Steven Abrams      11.00 United States          0
4  1983                 Luis Llamas      19.50        Mexico          0
5  1984               Birgit Felden       9.50       Germany          0
6  1985            Oscar Rodriguez      11.75 United States          0
7  1986                Mark Heller      15.50 United States          0
8  1987                Don Wolfman      12.00 United States          0
9  1988                  Jay Green      14.00 United States          0
10 1989                  Jay Green      13.00 United States          0
```

Рис. 3. Таблица с исходными данными, загруженная в R

Пробелы в названиях колонок были заменены точками. Вместо Dogs eaten теперь имеем Dogs.eaten. Для обращения к конкретной колонке данных вам нужно использовать название таблицы, за которым следует знак доллара $ и название колонки. Например, если вы хотите обратиться к Dogs.eaten, необходимо ввести следующее: hotdogs$Dogs.eaten

Теперь, когда данные уже в R, вы можете перейти к построению диаграммы командой barplot():

```
barplot(hotdogs$Dogs.eaten)
```

Вы отдаете распоряжение R начертить диаграмму на основе данных из колонки Dogs.eaten. В результате вы получите столбчатую диаграмму (рис. 4).

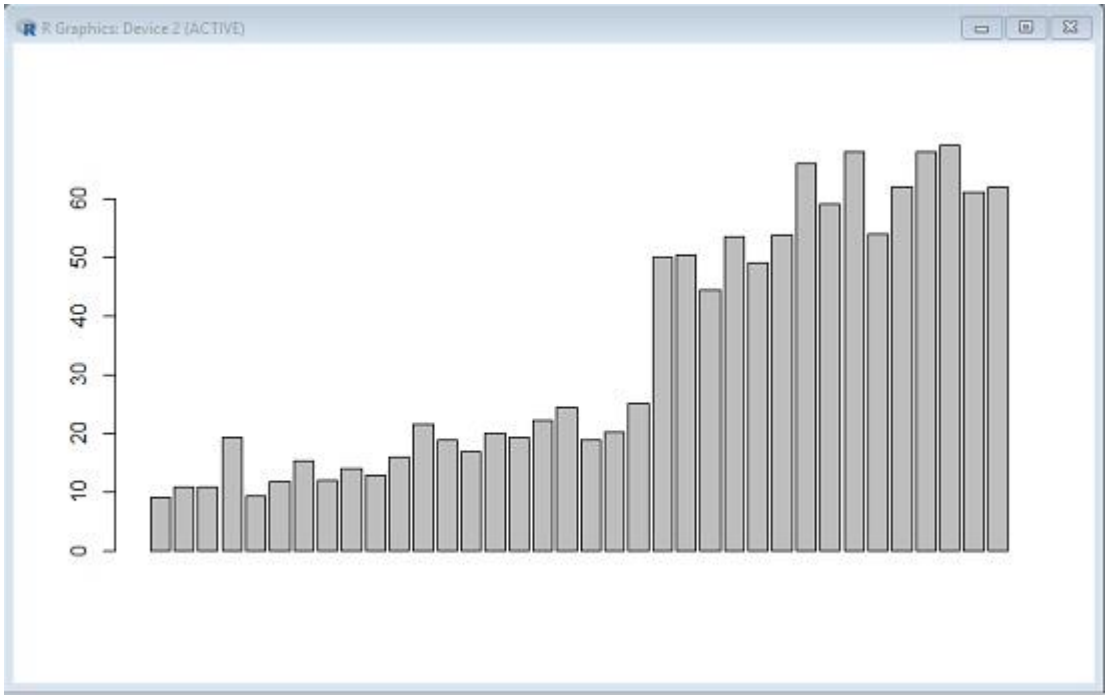


Рис. 4. Диаграмма по умолчанию, созданная с использованием barplot() в R, на которой представлено количество съеденных хот-догов

Используйте аргумент names.arg в barplot(), чтобы добавить значения по оси абсцисс – годы проведения соревнований (рис. 5):

barplot(hotdogs$Dogs.eaten, names.arg=hotdogs$Year)

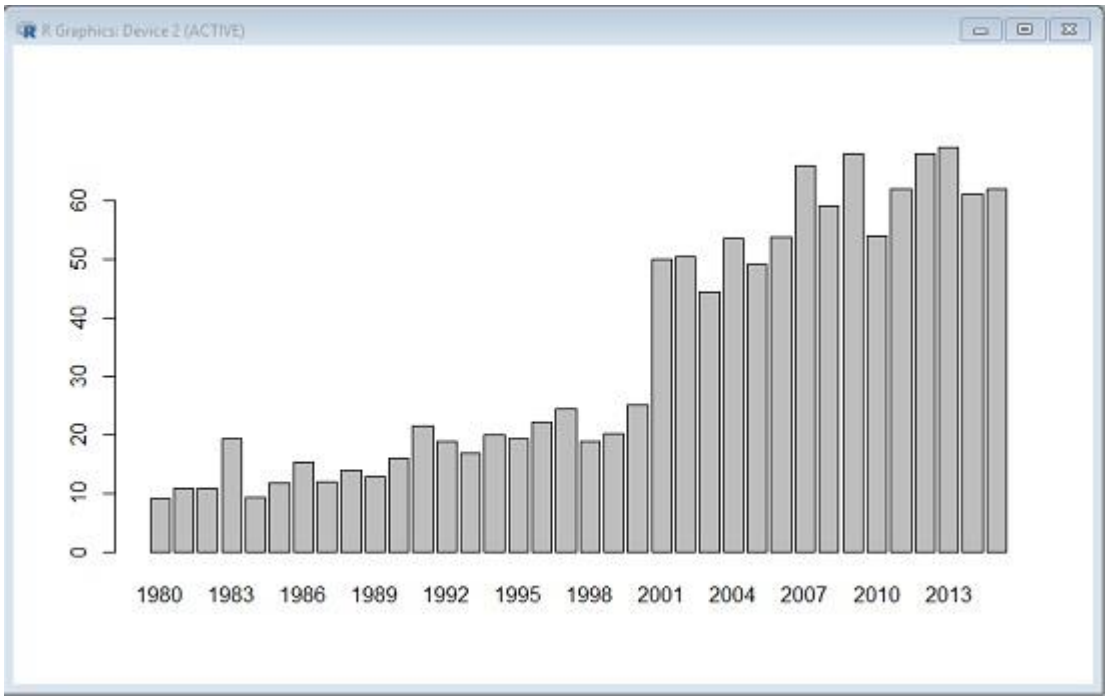


Рис. 5. Столбчатая диаграмма с обозначениями годов проведения соревнований

Можете разместить подписи к осям, изменить рамки и добавить цвета (рис. 6).

barplot(hotdogs$Dogs.eaten, names.arg=hotdogs$Year, col="red", border=NA, xlab="Годы", ylab="Число съеденных хот-догов")

Аргумент col служит для задания цвета (возможные варианты перечислены в документации к R) или шестнадцатеричного числа, такого как #821122. В данном случае был выбран вариант без границ — NA (это логическая константа, означающая «без значения»). Кроме того, были добавлены подписи к осям X и Y: «Годы» и «Число съеденных хот-догов».

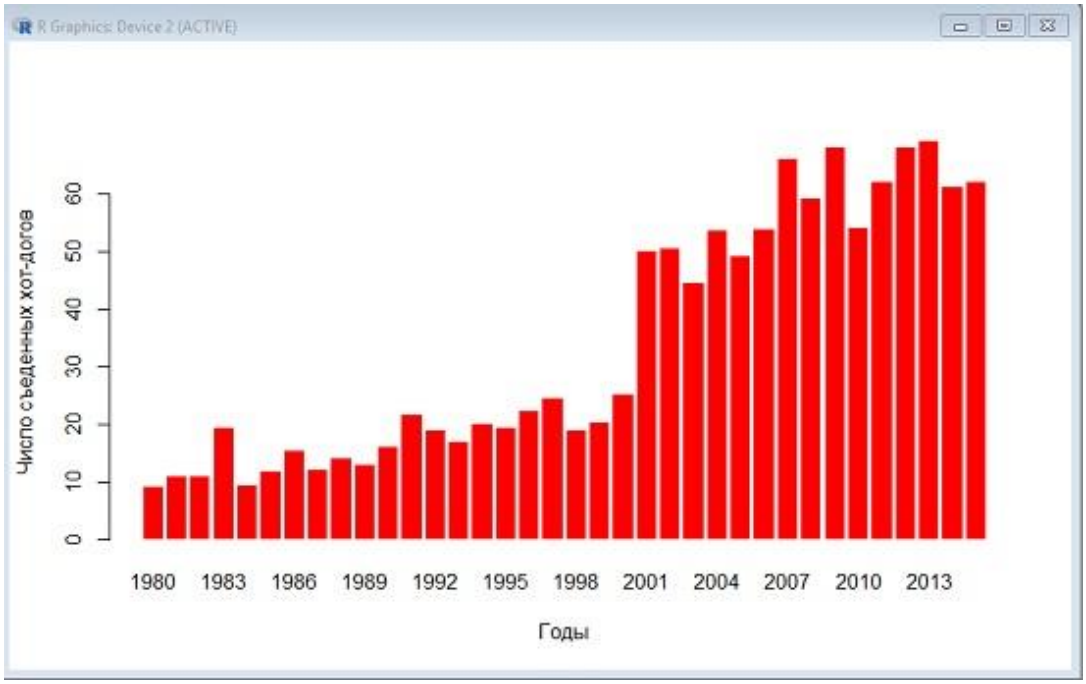


Рис. 6. Столбчатая диаграмма с окрашенными столбиками и подписями к осям

Вы можете перечислить в barplot() множество цветов, раскрасив каждый столбик, как хотите. Допустим, вам нужно выделить годы, когда Соединенные Штаты выигрывали соревнование. Эти годы вы можете окрасить в темно-красный цвет (#821122), а все остальные — в светло-серый (рис. 7).

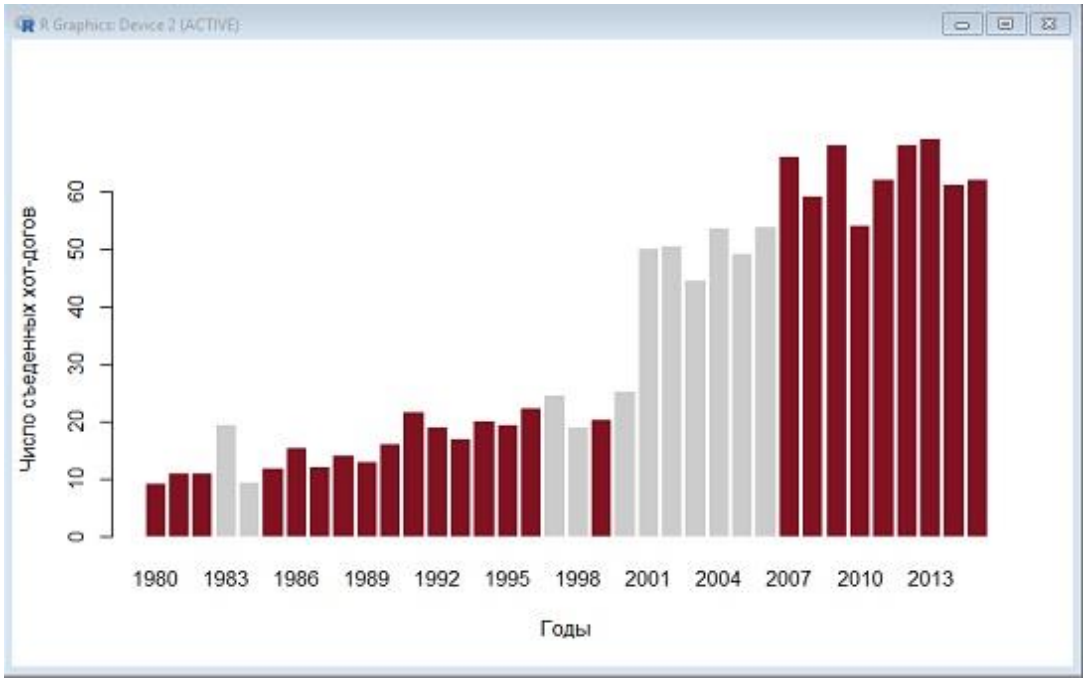


Рис. 7. Столбчатая диаграмма с индивидуально окрашенными столбиками

Чтобы сделать это, вам необходимо составить список (или вектор) цветов. Если выиграли США, пусть столбец будет темно-красным. В противном случае – серым:

```
# Highlight USA colors
fill_colors <- c()
for ( i in 1:length(hotdogs$Country) ) {
	if (hotdogs$Country[i] == "United States") {
		fill_colors <- c(fill_colors, "#821122")
	} else {
		fill_colors <- c(fill_colors, "#cccccc")
	}
}
```

Первая строка создает пустой вектор под именем fill_colors. В R для создания векторов используется команда c(). Следующая строка запускает цикл for для индекса i от 1 до числа строк в вашей таблице данных hotdogs. Вы можете взять колонку Country и вычислить ее длину (количество элемнетов). Если бы вы использовали length() только с hotdogs, тогда вы бы получили число колонок (5), но вас интересует количество строк (36). Для каждого года с 1980 по 2015 существует отдельная строка, а значит, цикл выполнит код внутри скобок 36 раз, и с каждым разом индекс i будет увеличиваться на 1.

При первой итерации, когда i равняется 1, производится проверка того, является ли страна в первом ряду (то есть страна победителя 1980 года) США. Если да, то с помощью fill_colors присваивается цвет #821122 (это и есть тот самый оттенок темно-красного в шестнадцатеричной форме). В противном случае присваивается #cccccc (серый). Чтобы увидеть, чем все закончится, введите в консоль fill_colors. Это тот самый вектор цветов, который вам нужен (рис. 8).

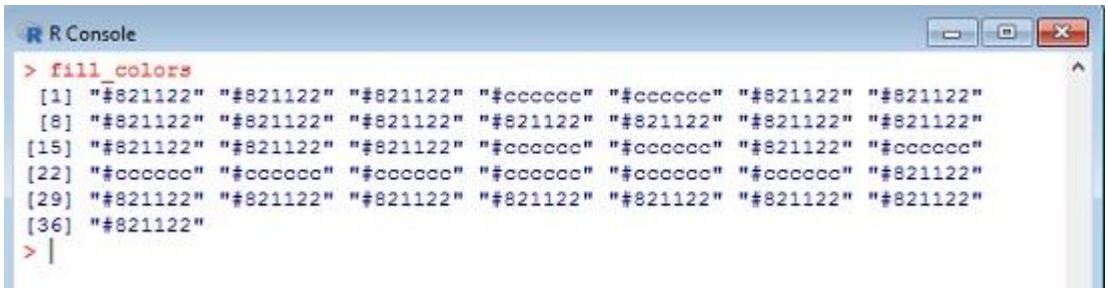


Рис. 8. Вектор цветов fill_colors

Передайте вектор fill_colors в аргумент col для barplot():

barplot(hotdogs$Dogs.eaten, names.arg=hotdogs$Year, col=fill_colors, border=NA, xlab="Годы", ylab="Число съеденных хот-догов")

В окончательном варианте столбчатой диаграммы (см. рис. 1) выделены, однако, не те годы, когда побеждали Соединенные Штаты, а те, когда устанавливался новый мировой рекорд. Информация о новых рекордах в таблице данных содержится в колонке New.record, и если там стоит 1, то столбец должен быть темно-красным, в противном случае — серым (рис. 9):

```
# Highlight record years with color
fill_colors <- c()
for ( i in 1:length(hotdogs$New.record) ) {
	if (hotdogs$New.record[i] == 1) {
		fill_colors <- c(fill_colors, "#821122")
	} else {
		fill_colors <- c(fill_colors, "#cccccc")
	}
}
barplot(hotdogs$Dogs.eaten, names.arg=hotdogs$Year, col=fill_colors, border=NA, xlab="Годы", ylab="Число съеденных хот-догов")
```

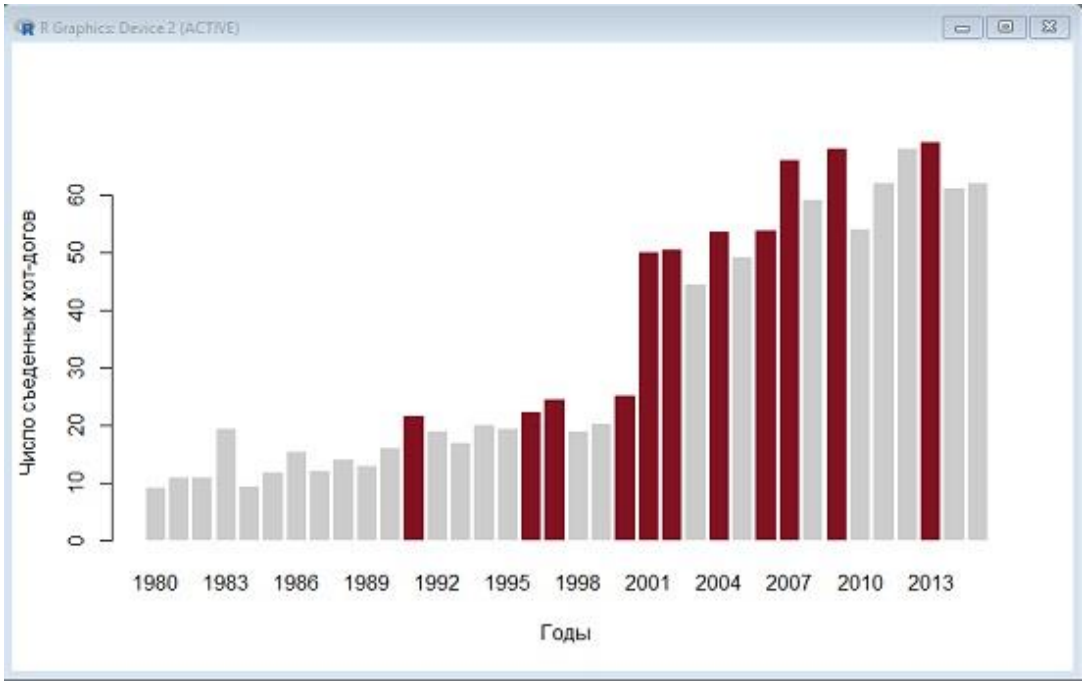


Рис. 9. Столбчатая диаграмма с подкрашенными рекордными значениями

На этом этапе вы можете поэкспериментировать и с другими аргументами barplot(), например изменить расстояние между столбцами или добавить заголовок (рис. 10). Обратите внимание, что интервал между столбцами стал больше, чем прежде, и появился заголовок.

barplot(hotdogs$Dogs.eaten, names.arg=hotdogs$Year, col=fill_colors, border=NA, space=0.3, xlab="Годы", ylab="Число съеденных хот-догов", main="Результаты Найтоновского турнира по поеданию хот-догов, 1980-2015")

Рис. 10. Столбчатая диаграмма с измененными интервалами между столбцами и заголовком

Можно сохранить графический файл в формате pdf. Для этого выделите окно графика, и пройдите в консоли по меню *Файл – > Сохранить как –> PDF*. Однако, если подписи содержат текст на русском языке, они будут отображаться некорректно. Сохраните файл с помощью следующего кода:

pdf("Nathan.pdf", family="NimbusSan")
barplot(hotdogs$Dogs.eaten, names.arg=hotdogs$Year, col=fill_colors, border=NA, space=0.5, xlab="Годы", ylab="Число съеденных хот-догов", main="Результаты Найтоновского турнира по поеданию хот-догов")

```
dev.off()
embedFonts("Nathan.pdf")
```

Чтобы просмотреть документацию по любой функции R, просто наберите знак вопроса и за ним название функции. Например, чтобы получить информацию о barplot(), просто наберите ?barplot. Вы перейдете на соответствующую страницу справки в Интернете. Справка обычно сопровождается примерами, которые бывают исключительно полезными.

Отредактируйте свою диаграмму в Illustrator. Посмотрите на диаграмму с точки зрения сторителлинга (подробнее см. Мартин Сайкс и др. От слайдов к историям. Пошаговая методика создания убеждающих презентаций). Представьте себе, что вы видите рис. 10 впервые. Вы — читатель и случайно наткнулись на эту диаграмму. Какую информацию вы можете почерпнуть из нее? Вы знаете, что на диаграмме отображено поедание хот-догов по годам. Значит ли это, что речь идет о пищевых привычках конкретного человека? Для одного человека хот-догов явно многовато. Может, это корм для животных? А остатки отдают птичкам? Или здесь представлено количество хот-догов, поедаемых людьми за год в среднем? И почему столбцы окрашены?

Поскольку вы — тот самый человек, который делал эту диаграмму, вам известен контекст, стоящий за каждым числом. Но читатели ничего не знают, а потому вам следует объяснить им, что здесь происходит. Хороший дизайн помогает читателям лучше вникнуть в историю. Программа Illustrator позволяет вручную манипулировать каждым из элементов графики. Вы можете задать шрифты, добавить примечания, видоизменить оси, отредактировать цвета и вообще сделать все, что подскажет вам ваша фантазия.

С сайта Adobe можно установить бесплатную пробную версию Illustrator (в качестве альтернативы можете попробовать Inkscape, распространяемую бесплатно). Откройте в Illustrator PDF-файл с вашей диаграммой. Обратите внимание на окно инструментов (рис. 11). Если вы не видите этого окна, то в пункте главного меню *Окно* выберите команду *Инструменты*.

Рис. 11. Окно инструментов Adobe Illustrator

Черная стрелка — это инструмент *Выделение*. Выберите его, и указатель вашей мыши превратится в черную стрелку. Щелкните ею по любому объекту. Появится выделение. В Illustrator это называется обтравочной маской. Она может пригодиться в самых разных ситуациях, но сейчас она вас не интересует, так что нажмите клавишу Delete на клавиатуре и избавьтесь от маски. Если в результате будет удалена вся диаграмма или ее элемент, отмените правку (Ctrl+Z) и используйте для выделения обтравочной маски инструмент *Прямое выделение*, который обозначен серой стрелкой.

Измените шрифт. Снова, используя инструмент *Выделение*, щелкните по тексту, чтобы отметить, что именно вы хотите изменить. Вызовите меню *Символ* (*Текст*), набрав Ctrl+T. Поменяйте шрифт на любой другой.

Выделите числовые подписи оси ординат. Перейдите в меню *Объект* и выберите *Трансформировать –> Трансформировать каждый*. Измените угол вращения на –90° (рис. 12). Подписи выстроятся по горизонтали (у меня команда работала не так как описано, и подписи выстраивались в одну горизонтальную линию, так что пришлось их растаскивать вручную). Поднимите подписи (только подписи, не линии разметки) вверх и вправо чтобы они оказались над метками на оси, а не слева от них.

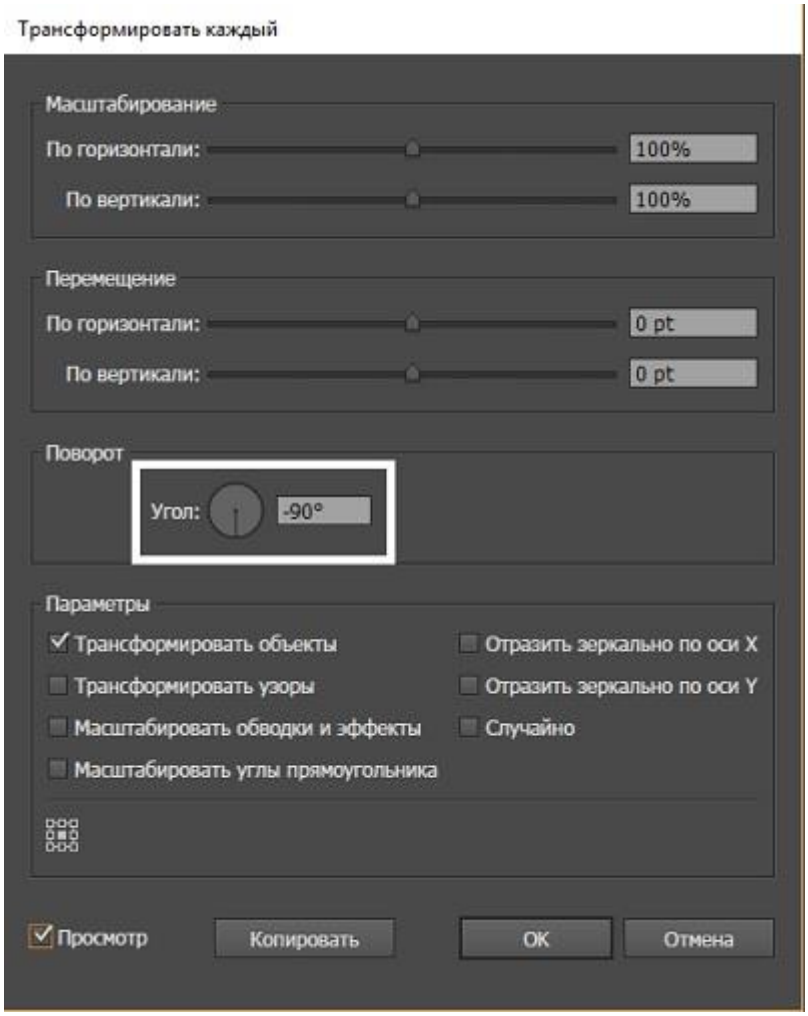


Рис. 12. Окно *Трансформировать каждый*

Удалите вертикальную линию оси значений. Она не несет никакой полезной информации. Если вы выберете вертикальную линию инструментом *Выделение*, подписи также окажутся выделенными. Так происходит, потому что они являются частью группы. Чтобы выбрать одну линию, используйте инструмент *Прямое выделение*. После того как линия окажется выделенной, нажмите клавишу Delete, и она исчезнет. Добавьте полужирную линию горизонтальной оси. Обратите внимание на то, что в окончательном варианте диаграммы мы имеем только штриховые метки (рис. 13). О том, какие элементы диаграммы только мешают восприятию см. [Принцип Эдварда Тафти минимизации количества элементов диаграммы].

Рис. 13. Ось значений преобразована

Существует множество способов создания линий разметки, но мы сейчас остановимся только на одном из них. Выберите инструмент *Перо* из окна инструментов, а затем установите стиль ваших линий в окне *Обводка*. Задайте толщину штриха 0,25 pt и убедитесь, что флажок *Пунктирная линия* не активен.

Чтобы начертить линию, щелкните мышью в том месте, где вы хотите расположить первую точку линии (или штриха разметки, как в данном случае), а затем щелкните там, где хотите поставить вторую точку. Если во время второго щелчка вы будете держать клавишу Shift, то автоматически получите прямую линию. Пока у вас есть одна метка. Но вам нужны еще 30 — по метке на каждый год. Вы можете проставить их все вручную, но есть способ лучше. Нажмите клавишу Alt. Таким образом будет создана копия, и у вас окажутся две метки. Далее нажмите Ctrl + D. Сделав это, вы продублируете новую метку, причем с таким же интервалом от предыдущей, как между первыми двумя. Нажимайте Ctrl + D до тех пор, пока не получите необходимое количество меток. Наконец, вам нужно расположить все метки правильно. Сдвиньте последнюю из них так, чтобы она оказалась в центре под последним столбцом. Первая метка и так уже должна находиться в центре под первым столбцом. Теперь выделите все метки и выберите в окне *Выравнивание* вариант *По горизонтали по центру* (рис. 14). После этого все метки распределятся на равных расстояниях между двумя крайними. Если вы захотите, то с помощью инструмента для *Выделение* можете выбрать каждую вторую метку и изменить ее размер по вертикали, сделав ее короче. Так с первого взгляда будет понятно, что более длинные метки относятся к сопроводительным подписям с годами. (Надо отметить, что и здесь мне не удалось воспользоваться рекомендациями автора, так что риски установил вручную, а затем их выровнял по вертикали.)

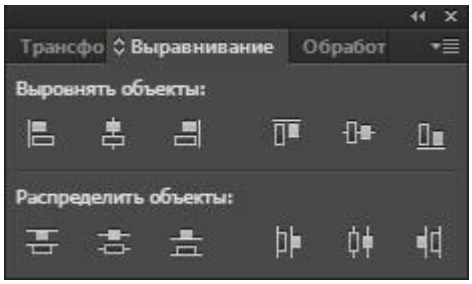


Рис. 14. Окно *Выравнивание*

Чтобы изменить цвет заливки с красного на зеленый, вы можете вернуться к инструменту *Прямого выделения* и выбрать по очереди все красные прямоугольники. Поскольку их относительно мало, это можно выполнить довольно быстро. Но что делать, если таких элементов много? Тогда лучше будет поступить следующим образом: щелкнуть по одному-единственному красному столбцу, а затем выбрать *Выделение –> По общему признаку –> С одинаковым цветом заливки*. Теперь поменяйте

цвет на тот, который вам больше нравится, через окно Цвет. Здесь вы можете перекрасить и границы, и заливку — однако сейчас вам нужно изменить только заливку (рис. 15).

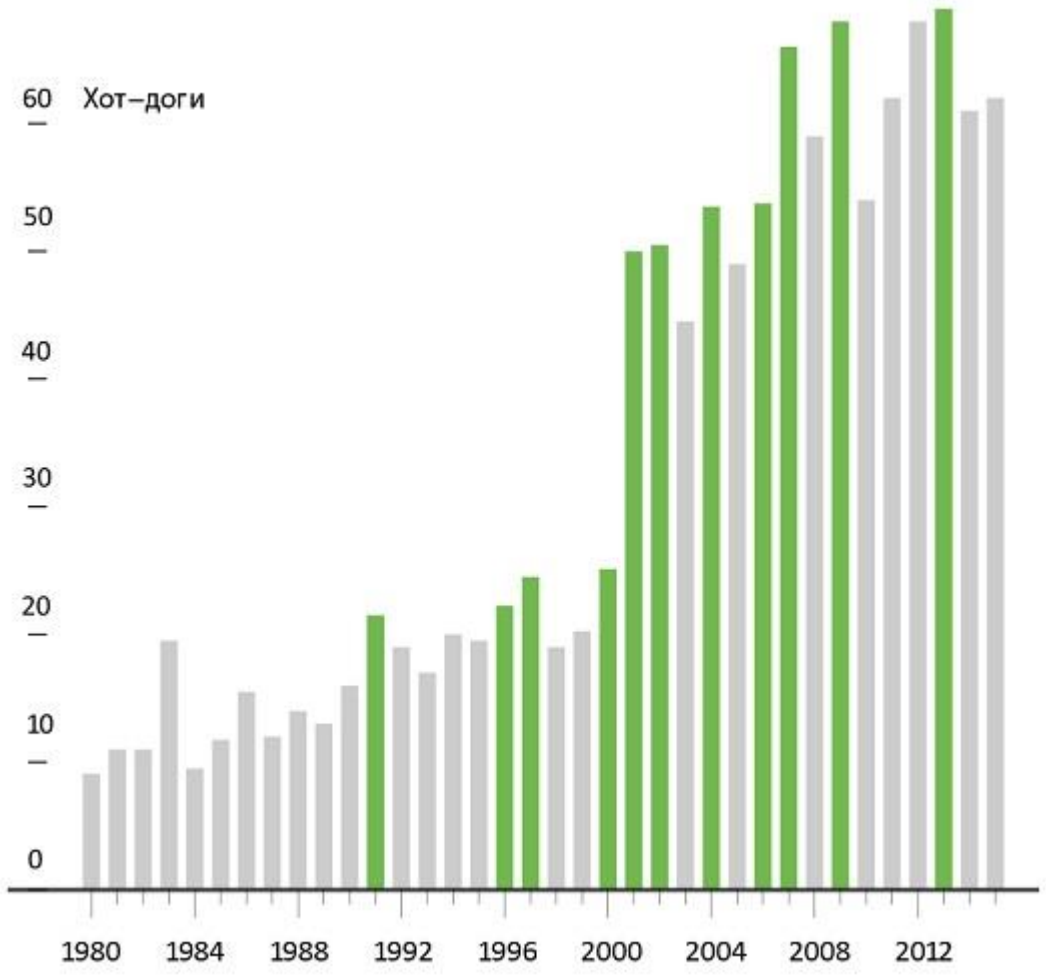


Рис. 15. Изменение цвета элементов диаграммы

Используйте инструмент *Текст*, который находится в окне *Инструменты*, чтобы добавить к диаграмме текстовые окошки. Это ваш шанс объяснить читателям, что именно они видят перед собой, глядя на диаграмму, и прояснить для них любые непонятные моменты. Выберите шрифт, который кажется вам подходящим, — причем установите другой кегль и гарнитуру, чтобы подписи отличались от остальных элементов диаграммы, таких как названия осей.

В случае с этой сверхважной хот-договой диаграммой особо стоит выделить первый с 1980 года рекорд, а также период, когда доминировал Такеру Кобаяши, а также сегодняшнее лидерство Джои Честната. Добавьте также заголовок и вводный абзац, объясняющий суть диаграммы. И вот что еще немаловажно: не забудьте указать источник данных. Без этого не будет никакого способа определить, насколько ваша диаграмма верна. Сложите все вместе, и вы получите окончательный вариант диаграммы, который представлен в самом начале.