Визуализация и анализ географических данных на языке R

Тимофей Самсонов

2017-09-26

Contents

Введение

Добро пожаловать в курс “Визуализация и анализ географических данных на языке R”! В данном курсе мы освоим азы программирования на языке R, а затем научимся использовать его для решения географических задач. Никаких предварительных знаний и навыков программирования не требуется.

Для успешного прохождения курса на вашем компьютере должно быть установлено следующее программное обеспечение:

- Язык R
- Среда разработки RStudio

Выбирайте инсталлятор, соответствующий вашей операционной системе. Обратите внимание на то, что RStudio не будет работать, пока вы не установите базовые библиотеки языка R. Поэтому обе вышеуказанные компоненты ПО обязательны для установки.

Установка и подключение пакетов

Существует множество дополнительных пакетов R (вы тоже можете написать свой) практически на все случаи жизни. Как и дистрибутив R, они доступны через CRAN (Comprehensive R Archive Network). Одним из таких пакетов является, например, пакет `openxlsx`, позволяющий читать и записывать файлы в форматах **Microsoft Excel**.

Существует два способа установки пакетов в **RStudio**.

Во-первых, вы можете сделать это в графическом интерфесе, нажав кнопку *Install* на панели *Packages* (по умолчанию эта панель расположена в нижней правой четверти окна программы). В появившемся окне введите название пакета и нажмите *Install*:

Во-вторых, вы можете вызвать *из консоли* команду `install.packages()`, передав ей в качестве параметра название пакета, заключенное в кавычки:

```
install.packages("openxlsx")
```

> Внимание: никогда не включайте команду `install.packages()` в тело скрипта. Это приведет к тому, что каждый раз при запуске программы среда **RStudio** будет пытаться заново установить пакет, который уже установлен. Запускайте эту функцию *только из консоли*.
>
> Если по каким-то причинам вы не можете установить пакет в стандартную системную директорию **RStudio** (например, из-за политик безопасности, запрещающих запись в каталог *Program Files* на ОС **Windows**), то необходимо создать директорию вручную в другом месте (куда вы имеете полный доступ) и указать ее адрес в параметре `lib` функции `install.packages()`. Например: `install.packages("xlsx", lib = "C:/Rlib/")`

Подключение пакета осуществляется с помощью функции `library()`, при этом название пакета можно в кавычки не заключать:

```
library(openxlsx)
```

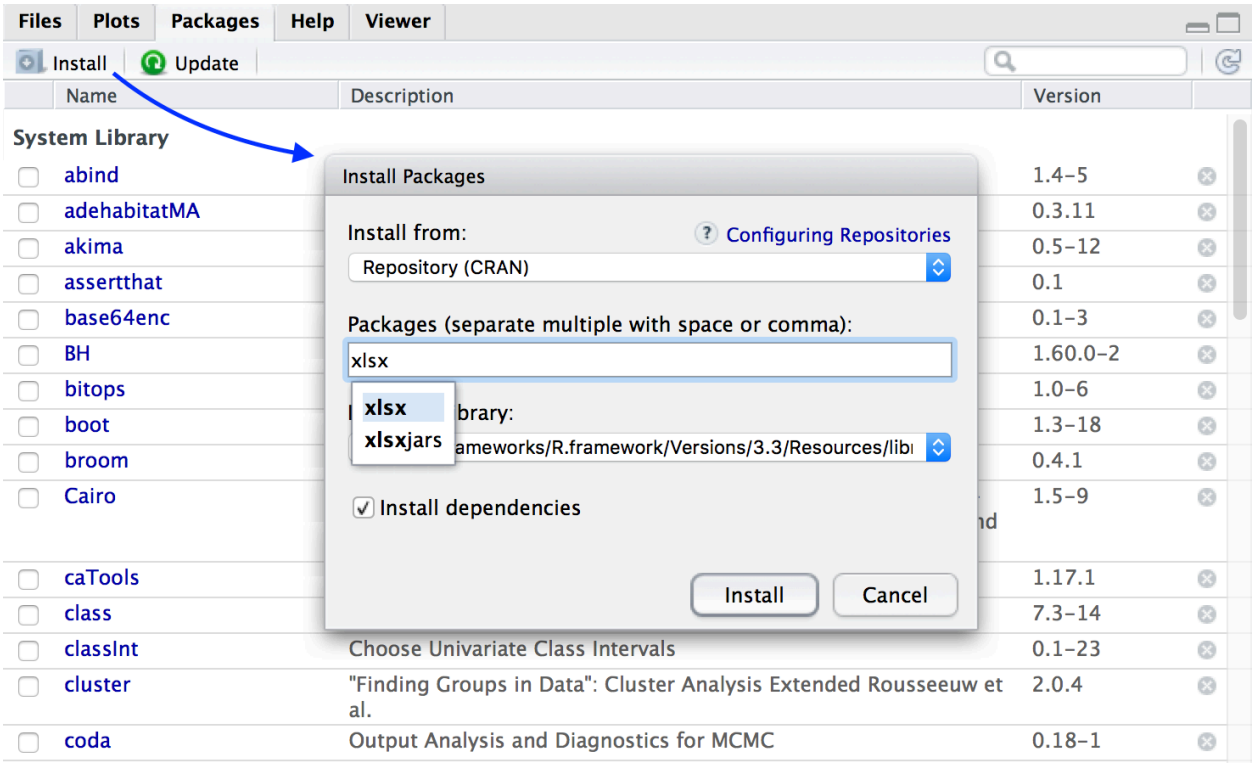

Figure 1: *Установка пакета*

> Если пакет установлен не в стандартный каталог, а в другое место — например, в каталог `:/Rlib/` (см. выше) — то при вызове функции `library()` необходимо указать местоположение пакета в дополнительном параметре `lib.loc`: `library(xlsx, lib.loc = "C:/Rlib")`

Форматы представления материалов

Прадлагаемый вашему вниманию курс состоит из ряда лекций (модулей), каждый из которых представлен в двух форматах: **HTML** и **R**.

Файл HTML

Файл **HTML** доступен в каждом модуле под названием *Лекция 1*, *Лекция 2* и т.д. Его удобно использовать в качестве справочника при выполнении заданий и подготовки к занятиям. Помимо текста и специальным образом форматированных фрагментов исходного кода, он содержит результаты генерации графиков и карт, а также навигацию по разделам.

Фрагменты исходного кода в тексте выглядят следующим образом:

```
a <- 5
sin(sqrt(a))
## [1] 0.7867491
sqrt(sin(a) + 2)
## [1] 1.020331
```

Обычным шрифтом `sin(sqrt(a))` в этих фрагментах показан исходный код, а бледным курсивом *`## [1]`*

0.7867491 — результат выполнения этого кода, который выводится в консоль среды разработки (в нашем случае это RStudio).

> Если вы хотите сохранить файл HTML себе и просматривать локально, достаточно после его открытия щелкнуть внутри страницы в любом месте правой кнопкой мыши (сокращенно ПКМ) и выбрать команду *Сохранить*, *Сохранить как* или *Сохранить фрейм как* (в зависимости от браузера формулировка может меняться).

Файл R

Файл **R** доступен в каждом модуле под названием *Исходный код лекции 1*, *Исходный код лекции 2* и так далее. Файл **R** также можно использовать при подготовке к лекциям и выполнении заданий. Содержание файла исходного кода полностью соответствует лекции. Отличие заключается в том, что файл **R** - это полноценно работающий скрипт. Все команды, приведенные в лекции, можно последовательно выполнить, а не только посмотреть на них, как это представлено в версии **HTML**. Исходный текст лекции в файле R преобразован в комментарии, так что вы можете последовательно читать и выполнять команды.

Весьма полезно также *создать копию* файла **R** и в процессе знакомства с новой темой попробовать поменять различные параметры и входные данные, чтобы посмотреть как меняется результат. В любом случае, даже если вы что-то испортите, исходный файл всегда доступен через систему электронного обучения, в которой вы в настоящий момент находитесь.

> Пользователи операционной системы Windows должны скачивать себе файл с суффиксом *CP1251* в названии, а OS X — с суффиксом *UTF8*. Отличие связано с тем, что в данных операционных системах используются разные кодировки для представления символов (текста). Если у вас при открытии файла отображаются кракозябры или знаки вопросов — это значит, что кодировка файла не соответствует вашей операционной системе или стандартным установкам RStudio. Проблему можно решить, скачав файл в нужной кодировке или воспользовавшись командой меню *File* > *Open With Encoding...*

Как работать с кодом

Существует несколько способов выполнения исходного кода:

- **Выполнить одну строку**: поставить курсор в любую строку и нажать над редактором кода кнопку *Run* или сочетание клавиш `Ctrl+Enter` (`Cmd+Enter` для OS X).
- **Выполнить несколько строк**: выделить необходимые строки и нажать над редактором кода кнопку *Run* или сочетание клавиш `Ctrl+Enter` (`Cmd+Enter` для OS X).
- **Выполнить весь код** можно сразу тремя способами:
  - Выделить весь текст и нажать над редактором кода кнопку *Run* или сочетание клавиш `Ctrl+Enter` (`Cmd+Enter` для OS X)
  - Нажать клавиатурное сочетание `Ctrl+Alt+Enter` (`Cmd+Alt+Enter` для OS X)
  - Нажать в правом верхнем углу редактора кода кнопку *Source*

    Команды *Source* и `Ctrl+Alt+Enter` могут не сработать, если у вас не установлена рабочая директория, или если в пути к рабочей директории содержатся кириллические символы (не актуально для Windows 10+ и OS X, которые являются системами, основанными на кодировке Unicode).

Существует также ряд дополнительных опций выполнения кода, которые вы можете найти в меню *Code* > *Run Region*

> Выполняя код построчно, делайте это последовательно, начиная с первой строки программы. Одна из самых распространенных ошибок новичков заключается в попытке выполнить некую строку, не выполнив *предыдущий код*. Нет никаких гарантий, что что-то получится, если открыть файл, поставить курсор в произвольную строку посередине программы и попытаться выполнить ее. Возможно, вам и повезет — если эта строка никак не зависит от предыдущего кода. Однако в реальных программах такие строки

составляют лишь небольшую долю от общего объема. Как правило, в них происходит инициализация новых переменных стартовыми значениями.

Стандарт оформления кода на R

Очень важно сразу же приучить себя грамотно, структурированно и красиво оформлять код на языке R. Это существенно облегчит чтение и понимание ваших программ не только вами, но и другими пользователями и разработчиками. Помимо вышеуказанных рекомендаций по написанию комментариев существует также определенное количество хорошо зарекомендовавших себя и широко используемых практик оформления кода. Эти практики есть в каждом языке программирования и их можно найти в литературе (и в Интернете) в виде негласных сводов правил (*style guides*)

Если вы не хотите быть белой вороной в мире R, вам будет полезно внимательно ознакомиться со стандартом оформления кода на R от компании Google, которая широко использует этот язык в своей работе.

> Стандарт оформления кода иногда также называют стилем программирования. Мы не будем использовать этот термин, поскольку под стилем программирования традиционно также понимают фундаментальный подход (*парадигму*) к построению программ: процедурный, функциональный, логический, объектно-ориентированный и некоторые другие.

К числу негласных правил оформления кода на R можно отнести следующие:

1. Последовательно используйте знак присвоения `<-` или `=` на протяжении всей программы. Если вы начали использовать `=` – применяйте его на протяжении всей программы, не используя `<-`.

   > Традиционный подход предполагает использование `<-`, однако все больше программистов использует знак `=` в своих программах, что делает R более похожим на другие языки программирования. Помните, что использование `=` официально не рекомендуется, поскольку существует много старого кода на R, который может ошибочно выполняться в сочетании с кодом, использующим `=`. Но вы, скорее всего, с такими проблемами не столкнетесь. Так что выбор за вами!

2. После запятой всегда ставьте пробел, перед запятой – нет:

```
#      :
a <- c(1, 2, 3, 4)
m <- matrix(a, 2, 2)

#       :
a <- c(1,2,3,4)
a <- c(1 ,2 ,3 ,4)
a <- c(1 , 2 , 3 , 4)
m <- matrix(a,2,2)
m <- matrix(a ,2 ,2)
m <- matrix(a , 2 , 2)
```

3. Отделяйте любые бинарные операторы (такие как `=`, `+`, `-`, `<-`, `*`) пробелами с двух сторон:

```
a <- sin(b + pi * 0.5) #
a<-sin(b+pi*0.5) #
```

4. Между названием функции и открывающей скобкой пробела быть не должно. То же самое касается обращения к элементам вектора, матрицы и т.п.:

```
#      :
sin(b)
a[2]
```

```
#       :
sin (b)
a [2]
```

5. В то же время, при вызове команд управления выполнением программы (условные операторы и циклы) перед и после скобок пробел **должен** стоять:

```
#      :
if (a > 0) {
  print(a)
}
i <- 0
while (i < a) {
  print(i)
  i <- i + 1
}

#       :
if(a > 0){
  print(a)
}

i <- 0
while(i < a){
  print(i)
  i <- i + 1
}
```

Комментарии

Комментарии — это фрагменты текста программы, начинающиеся с символа #. Комментарии не воспринимаются как исполняемый код и служат для документирования программы. При выполнении программы содержимое комментария в зависимости от настроек среды может выводиться или не выводиться в консоль, однако их содержание никак не влияет на результаты выполнения программы.

> Всегда пишите комментарии, чтобы по прошествии времени можно было открыть файл и быстро восстановить в памяти логику программы и смысл отдельных операций. Комментарии особенно необходимы, если вашей программой будет пользоваться кто-то другой — без них будет трудно разобраться в программном коде.

Действие комментария продолжается от символа # до конца строки. Соответственно, вы можете поставить данный символ в самом начале строки и тогда комментарий будет занимать всю строку. Комментарий также можно расположить справа от исполняемого кода, и тогда он будет занимать только часть строки.

> Прервать комментарий и написать справа от него исполняемый код нельзя

Полностроочные комментарии часто используются для выделения разделов в программе и написания объемных пояснений. Часто в них вводят имитации разделительных линий с помощью символов дефиса (−) или подчеркивания (_), а заголовки набирают прописными буквами. Короткие комментарии справа от фрагментов кода обычно служат пояснением конкретных простых операций. Подобная логика употребления комментариев не является обязательной. Вы можете оформлять их на свое усмотрение. Главное, чтобы они выполняли свою основную функцию — пояснять смысл выполняемых действий. Например:

```
#
# ---------------------------
```

```
#                                  ,          ,     ,     ,          ,                    :

a <- 3 + 2 #
b <- 4 ^ 8 #
c <- b %% a #

#
d <- c / a

#
e <- d * b
```

Однако, усердствовать с комментированием каждой мелочи в программе, разумеется, не стоит. Со временем у вас выработается взвешенный подход к документированию программ и понимание того, какие ее фрагменты требуют пояснения, а какие самоочевидны.

> Для быстрой вставки комментария, обозначающего новый раздел программы, воспользуйтесь командой меню *Code* > *Insert Section* или клавиатурным сочетанием `Ctrl+Shift+R` (`Cmd+Shift+R` для OS X)

Названия специальных символов

В **R**, как и во многих других языках программирования испльзуются различные специальные символы. Их смысл и значение мы узнаем по ходу изучения языка, а пока что выучите их названия, чтобы грамотно употреблять в своей речи

Символ	Название
$	доллар
#	шарп
&	амперсанд (решетка)
/	прямой слэш
\	обратный слэш
|	пайп (вертикальная черта)
^	циркумфлекс (крышечка)
@	эт (собачка)
~	тильда
''	одинарные кавычки
""	двойные кавычки
``	обратные кавычки

Как ссылаться

Если этот курс лекций оказался полезным для вас, и вы хотите процитировать его с списке литературы вашей работы, ссылку следует оформить как:

Самсонов Т.Е. **Визуализация и анализ географических данных на языке R.** М.: Географический факультет МГУ, 2017. DOI: 10.5281/zenodo.901911

Chapter 1

Атомарные типы данных

1.1 Числа

Числа — основной тип данных в R. К ним относятся *числа с плавающей точкой* и *целые числа*. В терминологии R такие данные называются *интервальными*, поскольку к ним применимо понятие интервала на числовой прямой. Целые числа относятся к *дискретным интервальным*, а числа с плавающей точкой — к *непрерывным интервальным*. Числа можно складывать, вычитать и умножать:

```
2 + 3
## [1] 5
2 - 3
## [1] -1
2 * 3
## [1] 6
```

Разделителем целой и дробной части является точка, а не запятая:

```
2.5 + 3.1
## [1] 5.6
```

Существует также специальный оператор для возведения в степень. Для этого вы можете использовать или двойной знак умножения (**) или *циркумфлекс* (^), который в обиходе называют просто “крышечкой”:

```
2 ^ 3
## [1] 8
2 ** 3
## [1] 8
```

Результат деления по умолчанию имеет тип с плавающей точкой:

```
5 / 3
## [1] 1.666667
5 / 2.5
## [1] 2
```

Если вы хотите чтобы деление производилось целочисленным образом (без дробной части) необходимо использовать оператор %/%:

```
5 %/% 3
## [1] 1
```

Остаток от деления можно получить с помощью оператора %%:

```
5 %% 3
## [1] 2
```

Вышеприведенные арифметические операции являются бинарными, то есть требуют наличия двух чисел. Числа называются “операндами”. Отделять операнды от оператора пробелом или нет — дело вкуса. Я предпочитаю отделять, так как это повышает читаемость кода. Следующие два выражения эквивалентны. Однако сравните простоту их восприятия:

```
5%/%3
## [1] 1
```

```
5 %/% 3
## [1] 1
```

Как правило, в настоящих программах числа в явном виде встречаются лишь иногда. Вместо этого для их обозначения используют переменные. В вышеприведенных выражениях мы неоднократно использовали число 3. Теперь представьте, что вы хотите проверить, каковы будут результаты, если вместо 3 использовать 4. Вам придется заменить все тройки на четверки. Если их много, то это будет утомительная работа, и вы наверняка что-то пропустите. Конечно, можно использовать поиск с автозаменой, но что если тройки надо заменить не везде? Одно и то же число может выполнять разные функции в разных выражениях. Чтобы избежать подобных проблем, в программе вводят переменные и присваивают им значения. Оператор присваивания значения выглядит как <-

```
a <- 5
b <- 3
```

Чтобы вывести значение переменной на экран, достаточно просто ввести его:

```
a
## [1] 5
b
## [1] 3
```

Мы можем выполнить над переменными все те же операции что и над константами:

```
a + b
## [1] 8
a - b
## [1] 2
a / b
## [1] 1.666667
a %/% b
## [1] 1
a %% b
## [1] 2
```

Легко меняем значение второй переменной с 3 на 4 и выполняем код заново.

```
b <- 4
a + b
## [1] 9
a - b
## [1] 1
a / b
## [1] 1.25
a %/% b
## [1] 1
a %% b
## [1] 1
```

Нам пришлось изменить значение переменной только один раз в момент ее создания, все последующие операции остались неизменны, но их результаты обновились!

Новую переменную можно создать на основе значений существующих переменных:

```
c <- b
d <- a+c
```

Посмотрим, что получилось:

```
c
## [1] 4
d
## [1] 9
```

Вы можете комбинировать переменные и заданные явным образом константы:

```
e <- d + 2.5
e
## [1] 11.5
```

Противоположное по знаку число получается добавлением унарного оператора − перед константой или переменной:

```
f <- -2
f
## [1] -2
f <- -e
f
## [1] -11.5
```

Операция взятия остатка от деления бывает полезной, например, когда мы хотим выяснить, является число четным или нет. Для этого достаточно взять остаток от деления на 2. Если число является четным, остаток будет равен нулю. В данном случае c равно 4, d равно 9:

```
c %% 2
## [1] 0
d %% 2
## [1] 1
```

1.1.1 Числовые функции

Прежде чем мы перейдем к рассмотрению прочих типов данных и структур данных нам необходимо познакомиться с функциями, поскольку они встречаются буквально на каждом шагу. Понятие функции идентично тому, к чему мы привыкли в математике. Например, функция может называться Z, и принимать 2 аргумента: `x` и `y`. В этом случае она записывается как `Z(x,y)`. Чтобы получить значение функции, необходимо подставить некоторые значения вместо `x` и `y` в скобках. Нас даже может не интересовать, как фактически устроена функция внутри, но важно понимать, что именно она должна вычислять. С созданием функций мы познакомимся позднее.

Важнейшие примеры функций — математические. Это функции взятия корня `sqrt(x)`, модуля `abs(x)`, а также тригонометрические функции `sin(x)`, `cos(x)`, `tan(x)` и обратные к ним `asin(y)`, `acos(y)`, `atan(y)` и так далее. В качестве аргумента функции можно использовать переменную, константу, а также выражения:

```
sqrt(a)
## [1] 2.236068
sin(a)
## [1] -0.9589243
tan(1.5)
## [1] 14.10142
```

```
abs(a + b - 2.5)
## [1] 6.5
```

Вы также можете легко вкладывать функции одна в одну, если результат вычисления одной функции нужно подставить в другую:

```
sin(sqrt(a))
## [1] 0.7867491
sqrt(sin(a) + 2)
## [1] 1.020331
```

Также как и с арифметическими выражениями, результат вычисления функции можно записать в переменную:

```
b <- sin(sqrt(a))
b
## [1] 0.7867491
```

Если переменной b ранее было присвоено другое значение, оно перезапишется. Вы также можете записать в переменную результат операции, выполненной над ней же. Например, если вы не уверены, что a — неотрицательное число, а вам это необходимо в дальнейших расчетах, вы можете применить к нему операцию взятия модуля:

```
b <- sin(a)
b
## [1] -0.9589243
b <- abs(b)
b
## [1] 0.9589243
```

1.2 Строки

Строки — также еще один важнейший тип данных. Строки состоят из символов. Чтобы создать строковую переменную, необходимо заключить текст строки в кавычки:

```
s <- "              ,             ( .     )"
s
## [1] "              ,             ( .     )"
```

Длину строки в символах можно узнать с помощью функции `nchar()`

```
nchar(s)
## [1] 56
```

Строки можно складывать так же как и числа. Эта операция называется *конкатенацией*. В результате конкатенации строки состыковываются друг с другом и получается одна строка. В отличие от чисел, конкатенация производится не оператором +, а специальной функцией `paste()`. Состыковываемые строки нужно перечислить через запятую, их число может быть произвольно

```
s1 <- "              ,"
s2 <- "           "
s3 <- "( .     )"
```

Посмотрим содержимое подстрок:

```
s1
## [1] "              ,"
s2
## [1] "           "
```

```
s3
## [1] "( .     )"
```

А теперь объединим их в одну:

```
s <- paste(s1, s2)
s
## [1] "              ,            "
s <- paste(s1, s2, s3)
s
## [1] "              ,             ( .     )"
```

Настоящая сила конкатенации проявляется когда вам необходимо объединить в одной строке некоторое текстовое описание (заранее известное) и значения переменных, которые у вас вычисляются в программе (заранее неизвестные). Предположим, вы нашли в программе что максимальная численность населения в Детройте пришлась на 1950 год и составила 1850 тыс. человек. Найденный год записан у вас в переменную `year`, а население в переменную `pop`. Вы их значения пока что не знаете, они вычислены по табличным данным в программе. Как вывести эту информацию на экран “человеческим”” образом? Для этого нужно использовать конкатенацию строк.

Условно запишем значения переменных, как будто мы их знаем

```
year <- 1950
pop <- 1850
```

```
s1 <- "                                  "
s2 <- "          "
s3 <- "  .   "
s <- paste(s1, year, s2, pop, s3)
s
## [1] "                                   1950           1850   .   "
```

Обратите внимание на то что мы конкатенировали строки с числами. Конвертация типов осуществилась автоматически. Помимо этого, функция сама вставила пробелы между строками.

1.3 Даты

Даты являются необходимыми при работе с временными данными. В географическом анализе подобные задачи возникают сплошь и рядом. Точность указания времени может быть самой различной. От года до долей секунды. Чаще всего используются даты, указанные с точностью до дня. Для создания даты используется функция `as.Date()`. В данном случае точка — это лишь часть названия функции, а не какой-то особый оператор. В качестве аргумента функции необходимо задать дату, записанную в виде строки. Запишем дату рождения автора (можете заменить ее на свою):

```
birth <- as.Date('1986/02/18')
birth
## [1] "1986-02-18"
```

Сегодняшнюю дату вы можете узнать с помощью специальной функции `Sys.Date()`:

```
current <- Sys.Date()
current
## [1] "2017-09-26"
```

Даты также можно складывать и вычитать. В зависимости от дискретности данных, вы получите результат в часах, днях, годах и т.д. Например, узнать продолжительность жизни в днях можно так:

```
livedays <- current - birth
livedays
## Time difference of 11543 days
```

Вы также можете прибавить к текущей дате некоторое значение. Например, необходимо узнать, какая дата будет через 40 дней:

```
current + 40
## [1] "2017-11-05"
```

С другими примерами использования дат мы познакомимся в дальнейшем по мере работы с данными.

1.4 Логические

Логические переменные возникают там, где нужно проверить условие. Переменная логического типа может принимать значение TRUE (истина) или FALSE (ложь). Для их обозначения также возможны более компактные константы T и F соответственно.

Следующие операторы приводят к возникновению логических переменных:

- *РАВНО* (==) — проверка равенства операндов
- *НЕ РАВНО* (!=) — проверка неравенства операндов
- *МЕНЬШЕ* (<) — первый аргумент меньше второго
- *МЕНЬШЕ ИЛИ РАВНО* (<=) — первый аргумент меньше или равен второму
- *БОЛЬШЕ* (>) — первый аргумент больше второго
- *БОЛЬШЕ ИЛИ РАВНО* (>=) — первый аргумент больше или равен второму

Посмотрим, как они работают:

```
a <- 1
b <- 2
a == b
## [1] FALSE
a != b
## [1] TRUE
a > b
## [1] FALSE
a < b
## [1] TRUE
```

Если необходимо проверить несколько условий одновременно, их можно комбинировать с помощью логических операторов. Наиболее популярные среди них:

- *И* (&&) - проверка истинности обоих условий
- *ИЛИ* (||) - проверка истинности хотя бы одного из условий
- *НЕ* (!) - отрицание операнда (истина меняется на ложь, ложь на истину)

```
c<-3
(b>a) && (c>b)
## [1] TRUE
(a>b) && (c>b)
## [1] FALSE
(a>b) || (c>b)
## [1] TRUE
!(a>b)
## [1] TRUE
```

Более подробно работу с логическими переменными мы разберем далее при знакомстве с условным оператором `if`.

1.5 Контрольные вопросы

Chapter 2

Векторы

В это модуле мы познакомимся с векторами – упорядоченными последовательностями объектов одного типа. Вектор является простейшей и одновременно базовой структурой данных в R. Понимание принципов работы с векторами необходимо для дальнейшего знакомства с более сложными структурами данных, такими как матрицы, фреймы данных, списки и массивы

2.1 Создание вектора

Вектор представляет собой упорядоченную последовательность объектов одного типа. То есть, вектор может состоять *только* из чисел, *только* из строк, *только* из дат или *только* из логических значений. Числовой вектор легко представить себе в виде набора цифр, выстроенных в ряд и пронумерованных согласно порядку их расстановки.

Существует множество способов создания векторов. Среди них наиболее употребительны:

1. Явное перечисление элементов
2. Создание пустого вектора (“болванки”), состоящего из заданного числа элементов
3. Генерация последовательности значений

Для создания вектора путем **перечисления** элементов используется функция `c()`:

```
#           -
colors <- c("    ", "    ", "   ", "      ", "     ")
colors
## [1] "    "    "    "    "   "      "      " "     "
```

```
#           -                 (                )
lengths <- c(28, 40, 45, 19, 38)
lengths
## [1] 28 40 45 19 38
```

```
#                   -                          (                )
opens <- c(FALSE, TRUE, TRUE, FALSE, FALSE)
opens
## [1] FALSE  TRUE  TRUE FALSE FALSE
```

> Внимание: не используйте латинскую букву ‘c’ в качестве названия переменной! Это приведет к конфликту названия встроенной функции c() и определенной вами переменной

Помимо этого, распространены сценарии, когда вам нужно создать вектор, но заполнять его значениями вы будете по ходу выполнения программы — скажем, при последовательной обработке строк таблицы. В этом случае вам известно

только предполагаемое количество элементов вектора и их тип. Здесь лучше всего подойдет **создание пустого вектора**, которое выполняется функцией `vector()`. Функция принимает 2 параметра:

- `mode` отвечает за тип данных и может принимать значения равные `"logical"`, `"integer"`, `"numeric"` (или `"double"`), `"complex"`, `"character"` и `"raw"`
- `length` отвечает за количество элементов

Например:

```
#       5      ,
intvalues <- vector(mode = "integer", length = 5)
intvalues #
## [1] 0 0 0 0 0

#       10      ,                                  (     )
charvalues <- vector("character", 10)
charvalues #
##  [1] "" "" "" "" "" "" "" "" "" ""
```

Обратите внимание на то, что в первом случае подстановка параметров произведена в виде = , а во втором указаны только значения. В данном примере оба способа эквивалентны. Однако первый способ безопаснее и понятнее. Если вы указываете только значения параметров, нужно помнить, что интерпретатор будет подставлять их именно в том порядке, в котором они перечислены в описании функции.

> Описание функции можно посмотреть, набрав ее название в консоли ее название со знаком вопроса в качестве префикса. Например, для вышеуказанной функции надо набрать `?vector`

Наконец, третий распространенный способ создания векторов — это **генерация последовательности**. Чтобы сформировать вектор из натуральных чисел от `M` до `N`, можно воспользоваться специальной конструкцией: `M:N`:

```
index <- 1:5 #        c(1,2,3,4,5)
index
## [1] 1 2 3 4 5
index <- 2:4 #        c(2,3,4)
index
## [1] 2 3 4
```

Существует и более общий способ создания последовательности — функция `seq()`, которая позволяет генерировать вектора значений нужной длины и/или с нужным шагом:

```
seq(from = 1, by = 2, length.out = 10) # 10         ,
##  [1]  1  3  5  7  9 11 13 15 17 19
seq(from = 2, to = 20, by = 3) #   2   20      3 (          )
## [1]  2  5  8 11 14 17 20
seq(length.out = 10, to = 2, by = -2) #                  10                    ,     2
##  [1] 20 18 16 14 12 10  8  6  4  2
```

Как видно, параметры функции `seq()` можно комбинировать различными способами и указывать в произвольном порядке (при условии, что вы используете полную форму (=). Главное, чтобы их совокупность *однозначно описывала последовательность*. Хотя, скажем, последний пример убывающей последовательности нельзя признать удачным с точки зрения наглядности.

Аналогичным образом можно создавать *последовательности дат*:

```
seq(from = as.Date('2016/09/01'), by = 1, length.out = 7) #                  2016/2017
## [1] "2016-09-01" "2016-09-02" "2016-09-03" "2016-09-04" "2016-09-05"
## [6] "2016-09-06" "2016-09-07"

seq(from = Sys.Date(), by = 7, length.out = 5) #             ,
```

```
## [1] "2017-09-26" "2017-10-03" "2017-10-10" "2017-10-17" "2017-10-24"
```

2.2 Работа с элементами вектора

К отдельным **элементам вектора** можно обращаться по их индексам:

```
colors[1] #
## [1] "    "
colors[3] #
## [1] "   "
```

Количество элементов (длину) вектора можно узнать с помощью функции `length()`:

```
length(colors)
## [1] 5
```

Последний элемент вектора можно извлечь, если мы знаем его длину:

```
n <- length(colors)
colors[n]
## [1] "     "
```

Последовательности удобно использовать для извлечения подвекторов. Предположим, нужно извлечь первые 4 элемента. Для этого запишем:

```
lengths[1:4]
## [1] 28 40 45 19
```

Индексирующий вектор можно создать заранее. Это удобно, если номера могут меняться в программе:

```
m <- 1
n <- 4
index <- m:n
lengths[index]
## [1] 28 40 45 19
```

Обратите внимание на то что по сути один вектор используется для извлечения элементов из другого вектора. Это означает, что мы можем использовать не только простые последовательности натуральных чисел, но и векторы из прозвольных индексов. Например:

```
index <- c(1, 3, 4) #          1, 3  4
lengths[index]
## [1] 28 45 19

index <- c(5, 1, 4, 2) #
lengths[index]
## [1] 38 28 19 40
```

2.3 Анализ и преобразования векторов

К числовым векторам можно применять множество функций. Прежде всего, нужно знать функции вычисления базовых параметров статистического ряда — минимум, максимум, среднее, медиана, дисперсия, размах вариации, среднеквадратическое отклонение, сумма:

```
min(lengths) #
## [1] 19
max(lengths) #
## [1] 45
range(lengths) #           =       -
## [1] 19 45
mean(lengths) #
## [1] 34
median(lengths) #
## [1] 38
var(lengths) #       (        -      , variation)
## [1] 108.5
sd(lengths) #                    (standard deviation)
## [1] 10.41633
sum(lengths) #
## [1] 170
```

Одной из мощнейших особенностей R является то что он не проводит различий между числами и векторами чисел. Поскольку R является матричным языком, каждое число представляется как вектор длиной 1 (или матрица $1 1$). Это означает, что любая математическая функция, применимая к числу, будет применима и к вектору:

```
lengths * 1000 #
## [1] 28000 40000 45000 19000 38000
sqrt(lengths) #
## [1] 5.291503 6.324555 6.708204 4.358899 6.164414

stations <- c(20, 21, 22, 12, 24) #

dens <- stations / lengths #                       =   -       /
dens
## [1] 0.7142857 0.5250000 0.4888889 0.6315789 0.6315789
```

2.4 Поиск и сортировка элементов

К важнейшим преобразованиям векторов относится их **сортировка**:

```
lengths2 <- sort(lengths) #
lengths2 #
## [1] 19 28 38 40 45
lengths #
## [1] 28 40 45 19 38

lengths2 <- sort(lengths, decreasing = TRUE) #                    .              decreasing
lengths2 #
## [1] 45 40 38 28 19
lengths #
## [1] 28 40 45 19 38
```

Другая распространенная задача — это **поиск индекса** элемента по его значению. Например, вы хотите узнать, какая ветка Московского метро (среди рассматриваемых) является самой длинной. Вы, конечно, легко найдете ее длину с помощью функции `max(lengths)`. Однако это не поможет вам узнать ее название, поскольку оно находится в другом векторе, и его индекс в массиве неизвестен. Поскольку векторы упорядочены одинаково, нам достаточно узнать, под каким индексом в массиве `lengths` располагается максимальный элемент, и затем извлечь цвет линии метро под тем

же самым индексом. Дл поиска индекса элемента используется функция `match()`:

```
l <- max(lengths) #
idx <- match(l, lengths) #              ,      l,       lengths
color <- colors[idx] #
color
## [1] "   "
```

Здесь непохо бы лишний раз потренироваться в конкатенации строк, чтобы вывести результат красиво!

```
s <- paste(color, "              -         .                ", l, " ")
s
## [1] "                  -         .                45   "
```

Ну и напоследок пример “матрешки”” из функций — как найти название самой плотной линии одним выражением:

```
colors[match(max(dens),dens)]
## [1] "    "
```

2.5 Контрольные вопросы

Chapter 3

Матрицы, фреймы данных и списки

В это модуле мы продвинемся дальше в изучении структур данных языка и рассмотрим такие важные его элементы как матрицы, фреймы данных и списки.

3.1 Матрицы

Матрица — это обобщение понятия вектора на 2 измерения. С точки зрения анализа данных матрицы ближе к реальным данным, посколько каждая матрица по сути представляет собой таблицу со столбцами и строками. Однако матрица, как и вектор, может содержать только элементы одного типа (числовые, строковые, логические и т.д.). Позже мы познакомимся с фреймами данных, которые не обладают подобным ограничением. А пока рассмотрим, как работать с двумерными данными на примере матриц.

Матрица, как правило, создается с помощью функции `matrix`, которая принимает 3 обязательных аргумента: вектор исходных значений, количество строк и количество столбцов:

```
v <- 1:12  #                          1   12
m <- matrix(v, nrow = 3, ncol = 4)
m
##      [,1] [,2] [,3] [,4]
## [1,]    1    4    7   10
## [2,]    2    5    8   11
## [3,]    3    6    9   12
```

По умолчанию матрица заполняется данными вектора по столбцам, что можно видеть в выводе программы. Если вы хотите заполнить ее по строкам, необходимо указать параметр `byrow = TRUE`:

```
m <- matrix(v, nrow = 3, ncol = 4, byrow = TRUE)
m
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    5    6    7    8
## [3,]    9   10   11   12
```

Доступ к элементам матрицы осуществляется аналогично вектору, за исключением того что нужно указать положение ячейки в строке и столбце:

```
m[2,4]  # 2    , 4
## [1] 8
m[3,1]  # 3    , 1
## [1] 9
```

Помимо этого, из матрицы можно легко извлечь одну строку или один столбец. Для этого достаточно указать только номер строки или столбца, а номер второго измерения пропустить до или после запятой. Результат является вектором:

```
m[2,]  # 2
## [1] 5 6 7 8
m[,3]  # 3 c
## [1]  3  7 11
```

К матрицам можно применять операции, аналогичные операциям над векторами:

```
log(m)  #
##          [,1]      [,2]     [,3]     [,4]
## [1,] 0.000000 0.6931472 1.098612 1.386294
## [2,] 1.609438 1.7917595 1.945910 2.079442
## [3,] 2.197225 2.3025851 2.397895 2.484907
sum(m)  #
## [1] 78
median(m) #
## [1] 6.5
```

А вот сортировка матрицы приведет к тому что будет возвращен обычный вектор:

```
sort(m)
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

К матрицам также применимы специальные функции, известные из линейной алгебры, такие как транспонирование и вычисление определителя:

```
t(m)  #
##      [,1] [,2] [,3]
## [1,]    1    5    9
## [2,]    2    6   10
## [3,]    3    7   11
## [4,]    4    8   12
m2<-matrix(-3:3,nrow = 3, ncol = 3)
## Warning in matrix(-3:3, nrow = 3, ncol = 3):         [7]
##                  [3]
m2
##      [,1] [,2] [,3]
## [1,]   -3    0    3
## [2,]   -2    1   -3
## [3,]   -1    2   -2
det(m2) #
## [1] -21
det(m)  #    !
## Error in determinant.matrix(x, logarithm = TRUE, ...): 'x'
```

Матрицы также можно перемножать с помощью специального оператора %*%. При этом, как мы помним, число столбцов в первой матрице должно равняться числу строк во второй:

```
m2 %*% m
##      [,1] [,2] [,3] [,4]
## [1,]   24   24   24   24
## [2,]  -24  -28  -32  -36
## [3,]   -9  -10  -11  -12
m %*% m2  #    !
## Error in m %*% m2:
```

Функция `match()`, которую мы использовали для поиска элементов в векторе, не работает для матриц. Вместо этого необходимо использовать функцию `which()`. Если мы хотим найти в матрице m позицию числа 8, то вызов функции будет выглядеть так:

```
which(m == 8, arr.ind = TRUE)
##      row col
## [1,]   2   4
```

В данном случае видно, что результат возвращен в виде матрицы 1×2. Обратите внимание на то, что колонки матрицы имеют названия. Попробуем использовать найденные индексы, чтобы извлечь искомый элемент:

```
indexes <- which(m == 8, arr.ind = TRUE)
row <- indexes[1,1]
col <- indexes[1,2]
m[row,col]
## [1] 8
```

Ура! Найденный элемент действительно равен 8.

Еще один полезный способ создания матрицы — это собрать ее из нескольких векторов, объединив их по строкам. Для этого можно использовать функции `cbind()` и `rbind()`. На предыдущем занятии мы создали векторы с длиной и количеством станций на разных ветках метро. Можно объединить их в одну матрицу:

```
lengths <- c(28, 40, 45, 19, 38)
stations <- c(20, 21, 22, 12, 24)
cbind(lengths, stations)  #
##      lengths stations
## [1,]      28       20
## [2,]      40       21
## [3,]      45       22
## [4,]      19       12
## [5,]      38       24
rbind(lengths, stations)  #
##          [,1] [,2] [,3] [,4] [,5]
## lengths    28   40   45   19   38
## stations   20   21   22   12   24
```

Строки и столбцы матрицы можно использовать как векторы при выполнении арифметических операций:

```
mm <- cbind(lengths, stations)
mm[,2]/mm[,1]  #              1
## [1] 0.7142857 0.5250000 0.4888889 0.6315789 0.6315789
```

Результат можно присоединить к уже созданной матрице:

```
dens <- mm[,2]/mm[,1]
mm<-cbind(mm, dens)
mm
##      lengths stations      dens
## [1,]      28       20 0.7142857
## [2,]      40       21 0.5250000
## [3,]      45       22 0.4888889
## [4,]      19       12 0.6315789
## [5,]      38       24 0.6315789
```

Содержимое матрицы можно просмотреть в более привычном табличном виде для этого откройте вкладку *Environment* и щелкните на строку с матрицей в разделе *Data*

Матрицы, однако, не дотягивают по функциональности до представления таблиц, и, в общем-то, не предназначены

для объединения разнородных данных в один набор (как мы это сделали). Если вы присоедините к матрице столбец с названиями веток метро, система не выдаст сообщение об ошибке, но преобразует матрицу в текстовую, так как текстовый тип данных способен представить любой другой тип данных:

```
colors <- c("    ", "    ", "   ", "      ", "     ")
mm2<-cbind(mm,colors)
mm2  #              ,
##      lengths stations dens                colors
## [1,] "28"    "20"     "0.714285714285714" "    "
## [2,] "40"    "21"     "0.525"             "    "
## [3,] "45"    "22"     "0.488888888888889" "   "
## [4,] "19"    "12"     "0.631578947368421" "      "
## [5,] "38"    "24"     "0.631578947368421" "     "
```

При попытке выполнить арифметическое выражение над прежде числовыми полями, вы получите сообщение об ошибке:

```
mm2[,2]/mm2[,1]
## Error in mm2[, 2]/mm2[, 1]:
```

3.2 Фреймы данных

Фреймы данных — это обобщение понятия матрицы на данные смешанных типов. Фреймы данных - наиболее распространенный формат представления табличных данных. Для краткости мы иногда будем называть их просто фреймами.

> Мы специально не используем для перевода слова `data.frame` термин ‘таблица’, поскольку таблица — это достаточно общая категория, которая описывает концептуальный способ упорядочивания данных. В том же языке R для представления таблиц могут быть использованы как минимум две структуры данных: фрейм данных (data.frame) и тиббл (tibble), доступный в соответствующем пакете. Мы не будем использовать тибблы в настоящем курсе, но после его освоения вы вполне сможете ознакомиться с ними самостоятельною

Для создания фреймов данных используется функция `data.frame()`:

```
t<-data.frame(colors,lengths,stations)
t  #        ,
##       colors lengths stations
## 1              28       20
## 2              40       21
## 3               45       22
## 4             19       12
## 5              38       24
```

К фреймам также можно пристыковывать новые столбцы:

```
t<-cbind(t, dens)
t
##       colors lengths stations      dens
## 1              28       20 0.7142857
## 2              40       21 0.5250000
## 3               45       22 0.4888889
## 4             19       12 0.6315789
## 5              38       24 0.6315789
```

Когда фрейм данных формируется посредством функции `data.frame()` и `cbind()`, названия столбцов берутся из

названий векторов. Обратите внимание на то, что листинге выше столбцы имеют заголовки, а строки — номера.

Как и прежде, к столбцам и строкам можно обращаться по индексам:

```
t[2,2]
## [1] 40
t[,3]
## [1] 20 21 22 12 24
t[4,]
##       colors lengths stations      dens
## 4            19       12 0.6315789
```

Вы можете обращаться к отдельным столбцам фрейма данных по их названию, используя оператор $ (доллар):

```
t$lengths
## [1] 28 40 45 19 38
t$stations
## [1] 20 21 22 12 24
```

Так же как и ранее, можно выполнять различные операции над столбцами:

```
max(t$stations)
## [1] 24
t$lengths / t$stations
## [1] 1.400000 1.904762 2.045455 1.583333 1.583333
```

Названия столбцов можно получить с помощью функции `colnames()`

```
colnames(t)
## [1] "colors"   "lengths"  "stations" "dens"
```

Чтобы присоединить строку, сначала можно создать фрейм данных из одной строки:

```
row<-data.frame("     ", 40.5, 22, 22/45)
```

Далее нужно убедиться, что столбцы в этом мини-фрейме называются также как и в той, куда мы хотим присоединить строку. Для этого нужно перезаписать результат, возвращаемый функцией `colnames()`:

```
colnames(row) <- colnames(t)
```

Обратите внимание на синтаксис вышеприведенного выражения. Когда функция возвращает результат, она обнаруживает свойство самого объекта, и мы можем его перезаписать. После того как столбцы приведены в соответствие, можно присоединить новую строку:

```
t<-rbind(t,row)
```

Поскольку названия столбцов хранятся как вектор из строк, мы можем их переделать:

```
colnames(t)<-c("  ","   ","    ","     ")
colnames(t)
## [1] "  "      "   "     "    "   "     "
```

Обратимся по новому названию столбца:

```
t$
## [1] 28.0 40.0 45.0 19.0 38.0 40.5
t
##
## 1          28.0      20 0.7142857
## 2          40.0      21 0.5250000
## 3           45.0      22 0.4888889
```

```
## 4         19.0       12 0.6315789
## 5         38.0       24 0.6315789
## 6         40.5       22 0.4888889
```

3.3 Списки

Список — это наиболее общий тип контейнера в R. Список отличается от вектора тем, что он может содержать набор объектов произвольного типа. В качестве элементов списка могут быть числа, строки, вектора, матрицы, фреймы данных — и все это в одном контейнере. Списки используются чтобы комбинировать разрозненную информацию. Результатом выполнения многих функций является список.

Например, можно создать список из текстового описания фрейма данных, самого фрейма данных и обобщающей статистики по нему:

```
d <- "                        6                "
s <- summary(t)   # summary()                        ,
```

Сооружаем список из трех элементов:

```
metrolist <- list(d,t,s)
metrolist
## [[1]]
## [1] "                        6                "
##
## [[2]]
##
## 1          28.0       20 0.7142857
## 2          40.0       21 0.5250000
## 3           45.0       22 0.4888889
## 4         19.0       12 0.6315789
## 5         38.0       24 0.6315789
## 6         40.5       22 0.4888889
##
## [[3]]
##
##         :1   Min.   :19.00   Min.   :12.00   Min.   :0.4889
##        :1   1st Qu.:30.50   1st Qu.:20.25   1st Qu.:0.4979
##         :1   Median :39.00   Median :21.50   Median :0.5783
##        :1   Mean   :35.08   Mean   :20.17   Mean   :0.5800
##          :1   3rd Qu.:40.38   3rd Qu.:22.00   3rd Qu.:0.6316
##        :1   Max.   :45.00   Max.   :24.00   Max.   :0.7143
```

Можно дать элементам списка осмысленные названия при создании:

```
metrolist <- list(desc = d, table = t, summary = s)
metrolist
## $desc
## [1] "                        6                "
##
## $table
##
## 1          28.0       20 0.7142857
## 2          40.0       21 0.5250000
## 3           45.0       22 0.4888889
```

```
## 4         19.0       12 0.6315789
## 5         38.0       24 0.6315789
## 6         40.5       22 0.4888889
##
## $summary
##
##         :1   Min.   :19.00   Min.   :12.00   Min.   :0.4889
##        :1   1st Qu.:30.50   1st Qu.:20.25   1st Qu.:0.4979
##         :1   Median :39.00   Median :21.50   Median :0.5783
##        :1   Mean   :35.08   Mean   :20.17   Mean   :0.5800
##          :1   3rd Qu.:40.38   3rd Qu.:22.00   3rd Qu.:0.6316
##        :1   Max.   :45.00   Max.   :24.00   Max.   :0.7143
```

Теперь можно обратиться к элементу списка по его названию:

```
metrolist$summary
##
##         :1   Min.   :19.00   Min.   :12.00   Min.   :0.4889
##        :1   1st Qu.:30.50   1st Qu.:20.25   1st Qu.:0.4979
##         :1   Median :39.00   Median :21.50   Median :0.5783
##        :1   Mean   :35.08   Mean   :20.17   Mean   :0.5800
##          :1   3rd Qu.:40.38   3rd Qu.:22.00   3rd Qu.:0.6316
##        :1   Max.   :45.00   Max.   :24.00   Max.   :0.7143
```

Поскольку `summary` сама является фреймом данных, из нее можно извлечь столбец:

```
metrolist$summary[,3]
##
## "Min.   :12.00  " "1st Qu.:20.25  " "Median :21.50  " "Mean   :20.17  "
##
## "3rd Qu.:22.00  " "Max.   :24.00  "
```

К элементу списка можно также обратиться по его порядковому номеру или названию, заключив их в *двойные* квадратные скобки:

```
metrolist[[1]]
## [1] "                      6              "
metrolist[["desc"]]
## [1] "                      6              "
```

Использование *двойных скобок* отличает списки от векторов.

3.4 Контрольные вопросы

Chapter 4

Чтение и обработка таблиц

Данный модуль посвящен введению в работу с таблицами. В модуле рассмотрены важные процедуры предварительной обработки таблиц, такие как фильтрация, исправление ошибок, преобразование типов — необходимые для дальнейшей визуализации и анализа данных

Выполните установку и подключение пакета `openxlsx`.

```
library(openxlsx)
```

4.1 Установка рабочей директории

Прежде чем мы начнем работать с данными, необходимо установить рабочую директорию. Это папка, в которой лежат необходимые вам таблицы. Для этого используем функцию `setwd()`. Аргумент функции нужно заменить на адрес каталога на вашем компьютере, в который вы положили присланные файлы:

```
setwd("/Volumes/Data/GitHub/r-geo-course/data")
```

> Внимание: если вы выполняете данный модуль на своем компьютере, замените вышеуказанный путь на путь к каталогу, в который вы положили исходные данные.

Теоретически рабочую директорию можно и не устанавливать, однако тогда при чтении файлов вам придется каждый раз указывать полный путь, что неудобно.

> Рабочая директория и местоположение скрипта могут не совпадать. Вы можете хранить их в разных местах. Однако рекомендуется держать их вместе, что облегчит передачу вашей программы вместе с данными другим пользователям.

4.2 Чтение таблиц CSV

Таблицы в формате **CSV** (Comma-Separated Values) можно прочесть с помощью универсальной функции `read.table()`. Следующие ее параметры важно указать:

- `file` — название файла
- `sep` — разделитель ячеек
- `dec` — десятичный разделитель
- `header` — содержится ли в первой строке заголовок — `encoding` — кодировка символов, в которой сохранен файл (чаще всего `UTF-8` или `CP1251`)

> Стандартной кодировкой для представления текста в UNIX-подобных системах (*Ubuntu*, *macOS* и т.д.) является **UTF-8 (Unicode)**, в русскоязычных версиях *Windows* — **CP1251 (Windows-1251)**. Текстовый файл **CSV**, созданный в разных операционных системах, будет по умолчанию сохраняться в соответствующей кодировке, если вы не указали ее явным образом. Если при загрузке таблицы в **R** вы видите вместо текста нечитаемые символы — *кракозябры* — то, скорее всего, вы читаете файл не в той кодировке, в которой он был сохранен. Попробуйте поменять `UTF-8` на `CP1251` или наоборот. Если вы не знаете, что такое кодировка и Юникод, то вам сюда.

Прочтем таблицу с данными Росстата по объему сброса сточных вод в бассейны некоторых морей России (в млн. м^3):

```
tab <- read.table("oxr_vod.csv",
                  sep = ';',
                  dec = ',',
                  header = TRUE,
                  encoding = 'UTF-8')
```

Для просмотра таблицы в привычном виде воспользуйтесь функцией `View()`. В этом представлении вы можете фильтровать и сортировать данные:

```
View(tab)
```

Show 5 entries Search:

	Год	Всего	Балтийское	Черное	Азовское	Каспийское	Карское	Белое	Прочие
1	1993	27.2	2.5	0.4	4.3	12.1	5.3	1	1.6
2	1994	24.6	2.3	0.4	3.2	11	5	0.9	1.8
3	1995	24.5	2.3	0.4	3.5	10.4	5.2	0.9	1.8
4	1996	22.4	2.2	0.3	3.1	9.8	4.7	0.8	1.5
5	1997	23	2.2	0.3	3.8	9.8	4.4	0.8	1.7

Showing 1 to 5 of 22 entries Previous 1 2 3 4 5 Next

Существуют более специальные функции для чтения таблиц **CSV**: `read.csv()` и `read.csv2()`. По сути они являются “обертками” (*wrappers*) функции `read.table()` и выполняют ее вызов с автоматической подстановкой параметров `sep`, `dec` и `header`. Обе функции по умолчанию предполагают, что в файле имеется заголовок. `read.csv()` удобна для чтения таблиц с десятичной точкой и запятой-разделителем, а `read.csv2()` — для таблиц с десятичной запятой и точкой-с-запятой в качестве разделителя.

Используем для чтения `read.csv2()`:

```
tab2<-read.csv2("oxr_vod.csv", encoding = 'UTF-8')
```

```
View(tab2)
```

Show 5 entries Search:

	Год	Всего	Балтийское	Черное	Азовское	Каспийское	Карское	Белое	Прочие
1	1993	27.2	2.5	0.4	4.3	12.1	5.3	1	1.6
2	1994	24.6	2.3	0.4	3.2	11	5	0.9	1.8
3	1995	24.5	2.3	0.4	3.5	10.4	5.2	0.9	1.8
4	1996	22.4	2.2	0.3	3.1	9.8	4.7	0.8	1.5
5	1997	23	2.2	0.3	3.8	9.8	4.4	0.8	1.7

Showing 1 to 5 of 22 entries Previous 1 2 3 4 5 Next

Как видно, данная таблица не отличается от предыдущей, но ее чтение более компактно.

4.3 Фильтрация, сортировка, работа с элементами таблицы

Распространенные операции с таблицами — это упорядочение по определенному столбцу и фильтрация по значениям. Мы уже знаем что из вектора, матрицы или таблицы можно извлекать элементы: `tab[V, ]`, где `tab` — имя таблицы, `V` — это вектор из номеров элементов. Например, извлечь 5, 2 и 4 строку таблицы можно так:

```
tab[c(5,2,4), ]
##
## 5 1997  23.0        2.2    0.3      3.8        9.8     4.4   0.8    1.7
## 2 1994  24.6        2.3    0.4      3.2       11.0     5.0   0.9    1.8
## 4 1996  22.4        2.2    0.3      3.1        9.8     4.7   0.8    1.5
```

Логично предположить, что таким же образом можно извлечь элементы таблицы в порядке, обеспечивающем возрастание или убывание значений в каком-то столбце. Для этого нужно правильным образом расставить индексы в векторе `c(...)`. Существует специальная функция `order()`, которая позволяет это сделать. Например, отсортируем таблицу по возрастанию сбросов в Каспийское море:

```
indexes<-order(tab$     )
head(tab[indexes, ])
##
## 22 2014  14.8        1.7    0.2      1.5        6.4     3.2   0.6    1.2
## 17 2009  15.9        1.8    0.2      1.5        6.8     3.5   0.7    1.4
## 21 2013  15.2        1.8    0.2      1.6        6.9     3.0   0.6    1.1
## 20 2012  15.7        1.8    0.2      1.6        7.0     3.0   0.7    1.4
## 19 2011  16.0        1.9    0.2      1.6        7.1     3.2   0.7    1.3
## 18 2010  16.5        2.0    0.2      1.6        7.3     3.3   0.7    1.4
```

> Используйте функцию `head()`, чтобы отобразить первые несколько строк таблицы. Эта возможность особенно полезна при работе с большими таблицами

Если упорядочение несложное, программист его скорее всего вставит непосредственно в инструкцию обращения к таблице:

```
head(tab[order(tab$     ), ])
##
## 22 2014  14.8        1.7    0.2      1.5        6.4     3.2   0.6    1.2
## 17 2009  15.9        1.8    0.2      1.5        6.8     3.5   0.7    1.4
## 21 2013  15.2        1.8    0.2      1.6        6.9     3.0   0.6    1.1
## 20 2012  15.7        1.8    0.2      1.6        7.0     3.0   0.7    1.4
## 19 2011  16.0        1.9    0.2      1.6        7.1     3.2   0.7    1.3
## 18 2010  16.5        2.0    0.2      1.6        7.3     3.3   0.7    1.4
```

Схожим образом реализована *фильтрация данных* по значению. Например, вы хотите извлечь из таблицы только те года, в которых объем сбросов в Каспийское море составил более 10 млн м^3. Здесь используется еще одна возможность извлечения элементов таблицы — с помощью вектора логических значений `TRUE/FALSE`. Число элементов в этом векторе должно быть равно числу элементов в индексируемом векторе, а значение указывает на то, нужно ли извлекать (`TRUE`) или нет (`FALSE`) элемент с текущим индексом. Вектор логических значений получается естественным путем с помощью операции сравнения:

```
condition <- tab$      > 10
condition  #
##  [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
tab[condition, ] #                          :
##
## 1 1993  27.2        2.5    0.4      4.3       12.1     5.3   1.0    1.6
## 2 1994  24.6        2.3    0.4      3.2       11.0     5.0   0.9    1.8
```

```
## 3 1995  24.5        2.3    0.4      3.5       10.4     5.2   0.9    1.8
```

Опять же, весьма часто используется запись одним выражением:

```
tab[tab$      > 10, ]
##
## 1 1993  27.2        2.5    0.4      4.3       12.1     5.3   1.0    1.6
## 2 1994  24.6        2.3    0.4      3.2       11.0     5.0   0.9    1.8
## 3 1995  24.5        2.3    0.4      3.5       10.4     5.2   0.9    1.8
```

Можно создать новую таблицу, выбрав необходимые столбцы:

```
caspian <- data.frame(tab$  , tab$   , tab$      )
```

Следует заметить, что не рекомендуется использовать кириллические названия столбцов (см. Правила подготовки таблиц для чтения в R в конце данного модуля), поэтому переименуем их:

```
colnames(caspian)<-c("Year", "Total", "Caspian")
```

Предположим, что теперь нам необходимо вычислить долю сбросов в Каспийское море в общем объеме и записать ее в новый столбец с точностью до 3 знаков после запятой. Для этого сначала произведем вычисления:

```
ratio <- caspian$Caspian / caspian$Total
ratio
##  [1] 0.4448529 0.4471545 0.4244898 0.4375000 0.4260870 0.4318182 0.4396135
##  [8] 0.4532020 0.4494949 0.4646465 0.4421053 0.4486486 0.4519774 0.4457143
## [15] 0.4302326 0.4385965 0.4276730 0.4424242 0.4437500 0.4458599 0.4539474
## [22] 0.4324324
```

Далее округлим результат с помощью функции `round()` с параметром `digits`, указывающим число значащих цифр в ответе:

```
ratio <- round(ratio, digits = 3)
ratio
##  [1] 0.445 0.447 0.424 0.438 0.426 0.432 0.440 0.453 0.449 0.465 0.442
## [12] 0.449 0.452 0.446 0.430 0.439 0.428 0.442 0.444 0.446 0.454 0.432
```

Существует простой и элегантный способ создать новый столбец в таблице — достаточно указать его название после значка $. Если среда R не обнаруживает столбец с таким названием, она его создаст:

```
caspian$CaspianRatio <- ratio
```

```
View(caspian)
```

Show 5 entries Search:

	Year	Total	Caspian	CaspianRatio
1	1993	27.2	12.1	0.445
2	1994	24.6	11	0.447
3	1995	24.5	10.4	0.424
4	1996	22.4	9.8	0.438
5	1997	23	9.8	0.426

Showing 1 to 5 of 22 entries Previous 1 2 3 4 5 Next

К столбцу таблицы можно обращаться по номеру, а не названию. Если вы указываете в квадратных скобках номер без запятой, он трактуется как номер столбца. При этом возвращаемый столбец имеет тип `data.frame`:

```
head(caspian[2])  #          (      -               )
##   Total
## 1  27.2
## 2  24.6
## 3  24.5
## 4  22.4
## 5  23.0
## 6  22.0
head(caspian[c(1,4)])  #
##   Year CaspianRatio
## 1 1993        0.445
## 2 1994        0.447
## 3 1995        0.424
## 4 1996        0.438
## 5 1997        0.426
## 6 1998        0.432
```

В противоположность этому, более привычная форма обращения к двумерным данным через запятую приведет к тому, что столбец будет возвращен как вектор:

```
caspian[,2]
##  [1] 27.2 24.6 24.5 22.4 23.0 22.0 20.7 20.3 19.8 19.8 19.0 18.5 17.7 17.5
## [15] 17.2 17.1 15.9 16.5 16.0 15.7 15.2 14.8
```

Использование той или иной формы зависит от контекста.

4.4 Чтение таблиц Microsoft Excel

Чтение таблиц **Microsoft Excel** производится с помощью функции `read.xlsx()` из пакета `openxlsx`. В качестве обязательных параметров они принимают следующие аргументы:

- `xlsxFile` — название файла
- `sheet` — номер листа

Убедитесь, что у вас установлена и подключена библиотека `openxlsx`.

Откроем таблицу с данными Росстата по сбросу загрязненных сточных вод в поверхностные водные объекты (млн м^3).

```
sewage<-read.xlsx("sewage.xlsx",1) #
## Warning in read.xlsx.default("sewage.xlsx", 1): '.Random.seed'
##              ,      'NULL',
```

```
View(sewage)
```

Show 5 entries Search:

	X1	2005	2010	2011	2012	2013
1	Российская Федерация	17727	16516	15966	15678	15189
2	Центральный федеральный округ	4341	3761	3613	3651	3570
3	Белгородская область	11	77	72	71	71
4	Брянская область	89	78	75	71	68
5	Владимирская область	155	129	126	124	120

Showing 1 to 5 of 97 entries Previous 1 2 3 4 5 … 20 Next

Следует дать адекватные названия столбцам таблицы:

```
colnames(sewage) <- c("Region", "Year05", "Year10", "Year11", "Year12", "Year13")
```

```
View(sewage)
```

Show 5 entries Search:

	Region	Year05	Year10	Year11	Year12	Year13
1	Российская Федерация	17727	16516	15966	15678	15189
2	Центральный федеральный округ	4341	3761	3613	3651	3570
3	Белгородская область	11	77	72	71	71
4	Брянская область	89	78	75	71	68
5	Владимирская область	155	129	126	124	120

Showing 1 to 5 of 97 entries Previous 1 2 3 4 5 … 20 Next

4.5 Пропущенные значения

Можно ли осуществлять обработку таблицы `sewage`? Попробуем в качестве примера найти минимум сбросов за 2012 год:

```
max(sewage$Year12)
## [1] NA
```

Результат имеет тип `NA`, потому что в данном столбце имеются пропуски. В некоторых статистических задачах это недопустимо. Если вы хотите проигнорировать значения пропусков, следует в вызываемой статистической функции указать дополнительный параметр `na.rm = TRUE`:

```
max(sewage$Year13, na.rm = TRUE)
## [1] 15189
```

Еще один вариант — исключить из таблицы те строки, в которых имеются пропущенные значения (хотя бы одно!). Для этого существует функция `complete.cases()`, возвращающая вектор логических значений:

```
filter<-complete.cases(sewage)
filter  #               .          FALSE -
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [12]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [23]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [34]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [45]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [56]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [67]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [78]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [89]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE

sewage.complete <- sewage[filter, ] #
```

```
View(sewage.complete)
```

Show 5 entries Search:

	Region	Year05	Year10	Year11	Year12	Year13
1	Российская Федерация	17727	16516	15966	15678	15189
2	Центральный федеральный округ	4341	3761	3613	3651	3570
3	Белгородская область	11	77	72	71	71
4	Брянская область	89	78	75	71	68
5	Владимирская область	155	129	126	124	120

Showing 1 to 5 of 93 entries Previous 1 2 3 4 5 … 19 Next

4.6 Фильтрация по текстовым полям

Часто бывает необходимо отобрать данные из таблицы, содержащей разнородные данные. В частности, в нашей таблице смешаны данные по субъектам и федеральным округам. Предположим, необходимо выгрузить в отдельную таблицу данные по федеральным округам. Для этого нужно найти строки, в которых столбец `Region` содержит фразу `"  "`. Для поиска по текстовым эталонам импользуется функция `grep()`, выдающая номера элементов, или ее разновидность `grepl()`, выдающая список логических констант

```
#          -          ,          -
rows <- grep("          ",sewage$Region)
rows  #      ,              Region
## [1]  2 21 35 42 49 64 73 86
okruga <- sewage[rows,] #
```

```
View(okruga)
```

Show 5 entries Search:

	Region	Year05	Year10	Year11	Year12	Year13
2	Центральный федеральный округ	4341	3761	3613	3651	3570
21	Северо-Западный федеральный округ	3192	3088	2866	2877	2796
35	Южный федеральный округ	1409	1446	1436	1394	1321
42	Северо-Кавказский федеральный округ	496	390	397	395	374
49	Приволжский федеральный округ	3162	2883	2857	2854	2849

Showing 1 to 5 of 8 entries Previous 1 2 Next

Наоборот — для **исключения** найденных объектов удобнее воспользоваться разновидностью `grepl()`, которая возвращает вектор из логических значений:

```
rows2 <- grepl("          ",sewage$Region)
rows2 #                grepl
##  [1] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [34] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [45] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [67] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
```

```
## [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

neokruga <- sewage[!rows2,]
```

```
View(neokruga)
```

Show 5 entries Search:

	Region	Year05	Year10	Year11	Year12	Year13
1	Российская Федерация	17727	16516	15966	15678	15189
3	Белгородская область	11	77	72	71	71
4	Брянская область	89	78	75	71	68
5	Владимирская область	155	129	126	124	120
6	Воронежская область	169	134	135	131	129

Showing 1 to 5 of 89 entries Previous 1 2 3 4 5 … 18 Next

Обратите внимание на восклицательный знак перед `rows2`. Он меняет все значения `TRUE` на `FALSE` и наоборот, что позволяет исключить найденные объекты

В полученной таблице все еще содержится текстовая шелуха типа " ", " . . . ", а также строка " ". К счастью, функция `grep()` достаточно умна и возволяет искать сразу по нескольким образцам строк. Для этого их нужно разделить вертикальной чертой — *пайпом* (`|`):

```
rows2 <- grepl("      |   |      | | ",sewage$Region)
rows2
##  [1]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## [23] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [34] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [45] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [67] FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
neokruga <- sewage[!rows2,] #                                rows2
```

```
View(neokruga)
```

Show 5 entries Search:

	Region	Year05	Year10	Year11	Year12	Year13
3	Белгородская область	11	77	72	71	71
4	Брянская область	89	78	75	71	68
5	Владимирская область	155	129	126	124	120
6	Воронежская область	169	134	135	131	129
7	Ивановская область	144	102	99	97	88

Showing 1 to 5 of 83 entries Previous 1 2 3 4 5 … 17 Next

4.7 Преобразование типов данных и исправление ошибок

Достаточно часто при работе с реальными данными возникает необходимость преобразования их типов. Например, вам необходимо перевести строки в даты, чтобы оперировать ими соответствующим образом. Или принудительным образом указать, что столбец со строками не хранит номинальную переменную (фактор), а его нужно интерпретировать именно как строковый столбец (обычно это полезно, когда столбец содержит какую-то текстовую информацию в виде комментариев по каждому измерению). Наконец, в данных могут быть ошибки, опечатки и так далее, которые могут препятствовать правильному их чтению.

В этом разделе мы рассмотрим, как можно:

1. Найти и исправить множественные варианты одного названия с опечатками
2. Исправить ошибки в числовых данных
3. Преобразовать факторы в строки и наоборот
4. Преобразовать строки в числа и наоборот

Рассмотрим возможные манипуляции с данными на примере таблицы о землепользовании на территории Сатинского учебного полигоны Географического факультета МГУ:

```
tab <- read.csv2("SatinoLanduse.csv", encoding = 'UTF-8')
str(tab) #      ,
## 'data.frame':    160 obs. of  6 variables:
##  $ ID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Type          : Factor w/ 12 levels "    ","    ",..: 11 1 1 9 5 1 5 12 9 10 ...
##  $ Administration: Factor w/ 7 levels "","  ","              ",..: 5 5 5 1 1 5 1 5 5 6 ...
##  $ Comment       : Factor w/ 35 levels "","  \"   \"",..: 30 1 1 1 1 1 1 3 2 1 ...
##  $ Perimeter     : Factor w/ 160 levels "1014.155593894044800",..: 67 155 51 104 78 153 17 19 108 57
##  $ Area          : Factor w/ 160 levels "0.238070145845919",..: 73 49 121 100 63 72 88 128 99 24 ...
```

```
View(tab)
```

Show 5 entries Search:

	ID	Type	Administration	Comment	Perimeter	Area
1	1	Территории населенных пунктов	Совьяковская сельская администрация	Село Беницы	2395.725463843722500	286159.158855028570000
2	2	Выгоны	Совьяковская сельская администрация		922.314163446381600	21651.963989321295000
3	3	Выгоны	Совьяковская сельская администрация		2180.787622893539300	56826.463220403260000
4	4	Пашни			3947.940266502537700	450293.758854912190000
5	5	Леса			278.569432215831970	2612.615315620203500

Showing 1 to 5 of 160 entries Previous 1 2 3 4 5 … 32 Next

Видно, что все столбцы, кроме двух, хранящих идентификаторы, были прочитаны как строки и преобразованы в факторы (номинальные переменные). Это означает, что мы не сможем работать привычным образом со столбцами периметра и площади, а столбец комментариев теперь также является номинальной переменной, что противоречит здравому смыслу (он вообще переменной не является).

Когда вы отображаете таблицу в консоли или графическом интерфейсе, факторы выглядят и ведут себя как обычные строки. Подвох заключается в том, что хранятся они в виде пар “ключ — значение” (об этом мы говорили выше) и все операции преобразования осуществляются **над ключами**, а не значениями. Рассмотрим, как следует правильно преобразовывать номинальные переменные в **R**.

Чтобы привести столбцы к нужному типу, необходимо использовать преобразования типов. Для этого в **R** существует множество функций семейства `as(object,class)`, где в качестве первого параметра `object` вы указываете преобразуемый объект, а в качестве второго параметра `class` — тип, к которому вы хотите его привести. Например:

```
s <- "5456.788"
s + 1
## Error in s + 1:
n <- as(s, "numeric")
## Error in as(s, "numeric"):              "as"
n + 1
## Error in eval(expr, envir, enclos):     'n'
s <- as(n, "character")
## Error in as(n, "character"):              "as"
s
## [1] "5456.788"
nchar(s)
## [1] 8
```

На практике обычно пользуются не функцией `as()`, а ее обертками (*wrappers*), которые имеют вид `as.numeric()`, `as.character()`, `as.Date()` и так далее:

```
as.numeric(s) #        ,      as(s, "numeric")
## [1] 5456.788
```

Для начала преобразуем столбец `Comment` к обычному символьному представлению:

```
tab$Comment <- as.character(tab$Comment)
str(tab)
## 'data.frame':    160 obs. of  6 variables:
##  $ ID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Type          : Factor w/ 12 levels "   ","    ",..: 11 1 1 9 5 1 5 12 9 10 ...
##  $ Administration: Factor w/ 7 levels "","," ",..: 5 5 5 1 1 5 1 5 5 6 ...
##  $ Comment       : chr  "       " "" "" "" ...
##  $ Perimeter     : Factor w/ 160 levels "1014.155593894044800",..: 67 155 51 104 78 153 17 19 108 57
##  $ Area          : Factor w/ 160 levels "0.238070145845919",..: 73 49 121 100 63 72 88 128 99 24 ...
```

Посмотрим теперь, что произойдет, если мы попытаемся преобразовать столбец `Perimeter` к числовому виду:

```
as.numeric(tab$Perimeter)
##   [1]  67 155  51 104  78 153  17  19 108  57   7   3 158 159 156  50  91
##  [18] 143   6  58   4   5 131 148 113 128 147 114   9  18 118 132  84 134
##  [35]  81  40 130  98  83 157  42  95  71 141   8 100  34   1  87  77 160
##  [52]  93 119  90  74  35 125 150 101 136  31 109 110 103  75  14  32  63
##  [69] 145  56 102  25  65  88  72  53  92  30 117  73  43  54 121  44  52
##  [86]  27 115 149 120  45  26  41   2  60  36 123  29 151 144 106 127  12
## [103] 116  94  82 146 142  69  21  48 139 105 154 124  47  61  33  80  97
## [120]  64  10  76 111  11 112  89  28 129  68  39  49  86  96  59  24 137
## [137]  46 152  55  15  99  85  22 126  16 122  79  66 133  23 107  38 138
## [154]  13 135  37 140  70  20  62
```

Вместо значений перметра мы получили загадочные числа, которых в таблице нет. Это и есть ключи факторов. Чтобы получить их значения, необходимо использовать функцию `levels()` (для краткости выведем первые 10 значений):

```
levels(tab$Perimeter)[1:10]
##  [1] "1014.155593894044800" "1019.457949256323400" "1020.278536197552200"
##  [4] "1021.109926202218700" "1041.122684298658400" "1060.678503301135200"
##  [7] "1081.964408568060900" "1094.945610298295600" "114.701418496307100"
## [10] "1155.916232728818800"
```

Обратите внимание на то, что значения фактора отсортированы в алфавитном порядке, без учете порядка их встречаемости в исходной таблице. Для корректного преобразования факторов в числа необходимо сначала

привести их к обычному строковому виду:

```
tab$Perimeter <- as.numeric(as.character(tab$Perimeter))
str(tab)
## 'data.frame':    160 obs. of  6 variables:
##  $ ID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Type          : Factor w/ 12 levels "   ","    ",..: 11 1 1 9 5 1 5 12 9 10 ...
##  $ Administration: Factor w/ 7 levels "","  ","             ",..: 5 5 5 1 1 5 1 5 5 6 ...
##  $ Comment       : chr  "       " "" "" "" ...
##  $ Perimeter     : num  2396 922 2181 3948 279 ...
##  $ Area          : Factor w/ 160 levels "0.238070145845919",..: 73 49 121 100 63 72 88 128 99 24 ...

#                       Area
temp <- as.numeric(as.character(tab$Area))
## Warning:                       NA
temp[1:10]
##  [1]  286159.159   21651.964   56826.463  450293.759    2612.615
##  [6]   28608.401 3469445.793   62299.631  450291.261  147943.134
```

Все прошло вроде бы успешно, но с предупреждением, что некоторые значения были преобразованы в `NA` (*Not Available*) — отсутствующие значения. По всей видимости, данные в соответствущих ячейках не соответствуют представлениям **R** о том, как должно выглядеть число: ячейка или пустая, или число набрано с ошибкой/опечаткой.

Чтобы найти и исправить все неверно заданные данные, необходимо выполнить следующие действия:

1. Получить индексы всех элементов, имеющих значение `NA`.
2. Просмотреть, какие значения были в исходных данных под этими индексами
3. Исправить ошибки в этих значениях, если это поддается автоматизации
4. Повторить конвертацию в числовой тип данных

Проверку на отсутствующие данные осуществляют с помощью функции `is.na()`. Передав ей в качестве аргумента вектор значений, вы получите вектор булевых значений, в котором `TRUE` будет стоять для пустых элементов. Проверим с помощью него, какие элементы столюца `Area` привели к ошибкам конвертации данных:

```
tab[is.na(temp), "Area"]
## [1] 89499,573298880117000 11922,638460079328000 5153,570673500797100
## 160 Levels: 0.238070145845919 ... 9865.323033935605100
```

Видно, что **R** не справился с преобразованием типов там, где содержится опечатка в десятичном разделителе — вместо точки указана запятая.

Для исправления этой ошибки мы можем воспользоваться стандартной функцией замены символа `gsub(pattern, replacement, x)`. Ее стандартные параметры означают соответственно: что искать, на что заменять, где искать:

```
tab$Area <- gsub(',', '.', tab$Area) #
tab$Area <- as.numeric(as.character(tab$Area)) #
str(tab)
## 'data.frame':    160 obs. of  6 variables:
##  $ ID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Type          : Factor w/ 12 levels "   ","    ",..: 11 1 1 9 5 1 5 12 9 10 ...
##  $ Administration: Factor w/ 7 levels "","  ","             ",..: 5 5 5 1 1 5 1 5 5 6 ...
##  $ Comment       : chr  "       " "" "" "" ...
##  $ Perimeter     : num  2396 922 2181 3948 279 ...
##  $ Area          : num  286159 21652 56826 450294 2613 ...
```

Теперь необходимо навести порядок в значениях факторов, убедившись, что и там нет опечаток. Выведем все уникальные значения с помощью функции `levels()`:

```
levels(tab$Type)
##  [1] "    "                          "     "
##  [3] "              "        "            "
##  [5] "   "                             "         "
##  [7] "   "                             "       "
##  [9] "    "                           "   "
## [11] "                  " "             "
levels(tab$Administration)
## [1] ""
## [2] " "
## [3] "                "
## [4] "                      "
## [5] "                     "
## [6] "                     "
## [7] "                    "
```

Видно, что если с типами все в порядке, то в данных об административном подчинении содержится 5 вариантов названия одной и той же Совьяковской сельской администрации. Помимо этого, пустые ячейки хорошо бы заменить на значение " ".

Чтобы найти все строчки, относящиеся к одному и тому же объекту, можно воспользоваться уже знакомой нам функцией `grep()`, передав ей подстроку, которая является для них общей. Например, " " (хотя в данном случае было бы вообще достаточно одной буквы " ").

```
filter <- grep("      ", tab$Administration) #
tab[filter, "Administration"] <- "                    " #
tab$Administration <- droplevels(tab$Administration) #
levels(tab$Administration)
## [1] ""
## [2] " "
## [3] "                    "
```

Пустые строки можно также найти с помощью `grep()`, но мы этого делать не будем, так как это требует дополнительных знаний о регулярных выражениях. Вместо этого воспользуемся тем, что пустые строки имеют длину 0. Обратите внимание ниже, что преобразование в вектор столбца `Administration` необходимо, т.к. `nchar()` не понимает объекты типа `data.frame`, которыми являются не только таблицы, но и их столбцы:

```
filter <- nchar(as.vector(tab$Administration)) == 0 # TRUE           0
#          :
tab[filter, "Administration"] <- "   "
## Warning in `[<-.factor`(`*tmp*`, iseq, value = c("   ", "   ", :
##                  ,        NA
```

Ошибка выше связана с тем, что **R** строго следит за неизменностью набора значений фактора для того чтобы избежать всевозможных ошибок при работе с данными (опечаток и т.д.). Предыдущий раз мы заменили все значаниея одним из существующих. В данном случае необходимо ввести новое значение фактора. Чтобы это сделать, придется преобразовать данные в символьные, произвести замену срок и после этого снова конвертировать столбец в фактор:

```
tab$Administration <- as.character(tab$Administration)
tab[filter, "Administration"] <- "   "
tab$Administration <- as.factor(tab$Administration)
levels(tab$Administration)
## [1] "   "
## [2] " "
## [3] "                    "
```

Теперь таблица готова к работе. Можно, например, подсчитать по ней сводную статистику:

```
summary(tab)
##        ID                                    Type
##  Min.   :  1.00                               :52
##  1st Qu.: 40.75                              :27
##  Median : 80.50                          :22
##  Mean   : 80.50                               :15
##  3rd Qu.:120.25                               :11
##  Max.   :160.00                     : 8
##                   (Other)                       :25
##                                Administration   Comment
##                                    :76     Length:160
##                                     : 3     Class :character
##                      :81     Mode  :character
##
##
##
##
##    Perimeter              Area
##  Min.   :    3.087   Min.   :      0
##  1st Qu.:  421.431   1st Qu.:   5087
##  Median :  939.369   Median :  21260
##  Mean   : 1761.654   Mean   : 125002
##  3rd Qu.: 2135.987   3rd Qu.:  83019
##  Max.   :23920.945   Max.   :3469446
##
```

Обратите внимание, что строки, интервальные и номинальные (факторы) переменные обрабатываются функцией `summary()` по-разному.

4.8 Сохранение таблиц CSV и Microsoft Excel

Одной из завершающих стадий анализа данных, помимо графиков и отчетов, часто являются новые табличные представления, которые было бы неплохо сохранить в виде файлов. К счастью, сохранение таблиц в **R** столь же просто, как и чтение. Для текстовых файлов в формате **CSV** можно использовать функции `write.table()`, `write.csv()` и `write.csv2()`. Для файлов **Microsoft Excel** используйте функцию `write.xlsx()` из пакета `openxlsx` соответственно.

> По умолчанию функции `write.table()`, `write.csv()` и `write.csv2()` записывают в таблицы в качестве первого столбца названия (номера) строк таблиц. Если вы не хотите, чтобы это происходило, укажите дополнительный параметр `row.names=FALSE`.

Сохраним таблицы `okruga` и `neokruga`, раздельно хранящие статистику по объему сброса сточных в поверхностные водные объекты по федеральным округам и субъектам соответственно:

```
write.csv2(okruga, "okruga.csv", fileEncoding = 'UTF-8') #                CSV        Unicode
write.xlsx(neokruga, "neokruga.xlsx") #                XLSX

#     ,                          :

okruga.saved <- read.csv2("okruga.csv", encoding = 'UTF-8')
head(okruga.saved)
##    X                              Region Year05 Year10 Year11 Year12
## 1  2                           4341   3761   3613   3651
## 2 21      -                   3192   3088   2866   2877
```

```
## 3 35                                1409   1446   1436   1394
## 4 42    -                         496    390    397    395
## 5 49                               3162   2883   2857   2854
## 6 64                                1681   1860   1834   1665
##   Year13
## 1   3570
## 2   2796
## 3   1321
## 4    374
## 5   2849
## 6   1624

neokruga.saved <- read.xlsx("neokruga.xlsx",1)
head(neokruga.saved)
##                 Region Year05 Year10 Year11 Year12 Year13
## 1                   11     77     72     71     71
## 2                    89     78     75     71     68
## 3                  155    129    126    124    120
## 4                  169    134    135    131    129
## 5                  144    102     99     97     88
## 6                   99     92     88     84     93
```

Видно, что в файле **CSV** присутствует также дополнительный столбец с названиями строк, а в файле **XLSX** его нет. Если вы не задавали названия строк явным образом и они не несут какого-то смысла, всегда указывайте параметр `row.names=FALSE`

> Вы можете дать строкам таблицы названия и извлечь их, используя функцию `row.names()` аналогично функции `colnames()` для столбцов.

4.9 Правила подготовки таблиц для чтения в R

С таблицами, которые мы использовали в настоящем модуле, все прошло гладко, поскольку они были подготовлены специальным образом. Несмотря на то, что каких-то четких правил подготовки таблиц для программной обработки не существует, можно дать несколько полезных рекомендаций по данному поводу:

1. В первой строке таблицы должны располагаться названия столбцов.
2. В названиях столбцов недопустимы объединенные ячейки, покрывающие несколько столбцов. Это может привести к неверному подсчету количества столбцов и, как следствие, некорректному чтению таблицы в целом.
3. Названия столбцов должны состоять из латинских букв и цифр, начинаться с буквы и не содержать пробелов. Сложносочиненные названия выделяйте прописными буквами. Плохое название столбца: `2015 ..` Хорошее название столбца: `GDP2015`.
4. Во второй строке таблицы должны начинаться данные. Не допускайте многострочных заголовков.
5. Некоторые ошибки данных в таблицах (такие как неверные десятичные разделители), проще найти и исправить в табличном/текстовом редакторе, нежели после загрузки в R.

Следование этим правилам значительно облегчит работу с табличными данными.

4.10 Контрольные вопросы

Bibliography