Data manipulation

Contents

```r
library(ISLR)
library(tidyverse)
library(haven)
```

Data types

There are several data types in R. Here is a table with the most common ones:

Type	Short	Example
Integer	int	0, 1, 2, 3, -4, -5
Numeric / Double	dbl	0.1, -2.5, 123.456
Character	chr	"dav is a cool course"
Logical	lgl	TRUE / FALSE
Factor	fct	low, medium, high

The `class()` function can give you an idea about what type of data each variable contains.

---

1. **Run the following code in R and inspect their data types using the `class()` function. Try to guess beforehand what their types will be!**

```r
object_1 <- 1:5
object_2 <- 1L:5L
object_3 <- "-123.456"
object_4 <- as.numeric(object_2)
object_5 <- letters[object_1]
object_6 <- as.factor(rep(object_5, 2))
object_7 <- c(1, 2, 3, "4", "5", "6")
```

the factor data type is special to R and uncommon in other programming languages. It is used to represent categorical variables with fixed possible values. For example, when there is a multiple choice question with 5 possible choices (a to e) and 10 students answer the question, we may get a result as in `object_6`.

Vectors can have only a single data type. Note that the first three elements in `object_7` have been converted. We can convert to different data types using the `as.<class>()` functions.

---

2. **Convert `object_7` back to a vector of numbers using the `as.numeric()` function**

```
object_7 <- as.numeric(object_7)
```

Lists

A list is a collection of objects. The elements may have names, but it is not necessary. Each element of a list can have a different data type, unlike vectors.

---

3. **Make a list called `objects` containing object 1 to 7 using the `list()` function.**

```
objects <- list(object_1, object_2, object_3, object_4, object_5, object_6,
                object_7)
```

A special type of list is the `data.frame`. It is the same as a list, but each element is forced to have the same length. The elements of a `data.frame` are the columns of a dataset. In the tidyverse, `data.frames` are called `tibbles`.

---

4. **Make a data frame out of `object_1`, `object_2`, and `object_5` using the `data.frame()` function**

```
dat <- data.frame(Var1 = object_1, Var2 = object_2, Var3 =object_5)
```

Loading data

We are going to use a dataset from Kaggle - the Google play store apps data by user `lava18`. We have downloaded it into the data folder already from https://www.kaggle.com/lava18/google-play-store-apps (downloaded on 2018-09-28).

Tidyverse contains many data loading functions – each for their own file type – in the packages `readr` (default file types) and `haven` (external file types such as from SPSS or Stata). The most common file type is `csv`, which is what we use here.

---

1. **Use the function `read_csv()` to import the file "data/googleplaystore.csv" and store it in a variable called `apps`.**

```r
apps <- read_csv("data/googleplaystore.csv")
```

```
## Parsed with column specification:
## cols(
##   App = col_character(),
##   Category = col_character(),
##   Rating = col_double(),
##   Reviews = col_integer(),
##   Size = col_character(),
##   Installs = col_character(),
##   Type = col_character(),
##   Price = col_character(),
##   `Content Rating` = col_character(),
##   Genres = col_character(),
##   `Last Updated` = col_character(),
##   `Current Ver` = col_character(),
##   `Android Ver` = col_character()
## )
```

If necessary, use the help files. These import functions from the tidyverse are fast and safe: they display informative errors if anything goes wrong. `read_csv()` also displays a message with information on how each column is imported: which variable type each column gets.

---

2. **Did any column get a variable type you did not expect?**

```r
# Several columns such as price and number of installs were imported as
# character data types, but they are numbers.
```

---

3. **Use the function `head()` to look at the first few rows of the `apps` dataset**

```r
head(apps)
```

```
## # A tibble: 6 x 13
##   App   Category Rating Reviews Size  Installs Type  Price `Content Rating`
##   <chr> <chr>     <dbl>   <int> <chr> <chr>    <chr> <chr> <chr>
## 1 Phot~ ART_AND~    4.1     159 19M   10,000+  Free  0     Everyone
## 2 Colo~ ART_AND~    3.9     967 14M   500,000+ Free  0     Everyone
## 3 "U L~ ART_AND~    4.7   87510 8.7M  5,000,0~ Free  0     Everyone
## 4 Sket~ ART_AND~    4.5  215644 25M   50,000,~ Free  0     Teen
## 5 Pixe~ ART_AND~    4.3     967 2.8M  100,000+ Free  0     Everyone
## 6 Pape~ ART_AND~    4.4     167 5.6M  50,000+  Free  0     Everyone
## # ... with 4 more variables: Genres <chr>, `Last Updated` <chr>, `Current
## #   Ver` <chr>, `Android Ver` <chr>
```