R Lesson 3: Creating Simple Scripts

vanderbi.lt/r

Steve Baskauf

Analysis of continuous data

DIGITAL SCHOLARSHIP AND COMMUNICATIONS

The plot() function

- The plot() function can graph two variables

```
plot(y ~ x)
```

- The dependent variable is listed before the tilde
- The independent variable is listed after the tilde

Linear model function

- The linear model function is used in R to analyze the relationship between two variables

```
model <- lm(y ~ x)
```

- As with plot, the independent variable is before the tilde and the dependent after
- By itself, **`lm()`** doesn't do much, but:

```
abline(model)
summary(model)
anova(model)
```

Schools data example

Factors

Factors

- A **factor** is a data structure for categorizing data.
- Its origin comes from **experimental design** terminology.
- In an experiment, each **category** into which an experimental trial can fall is called a **level**.
- Factors are sometimes called **grouping variables** because they are used to group observations.
- Factors may be required for some statistical tests and visualizations.

Factor example: science fair

water factor	height (cm)
wet	25
wet	21
dry	14
wet	13
dry	10
wet	18

- The water factor has two levels: wet and dry
- The height observations can be grouped by whether the experimental treatment was wet or dry

Factor example: creating factor values

- Create a vector of character strings and a vector of number values:

```
water_conditions <- c("wet", "wet", "dry", "wet", "dry", "wet")
height <- c(25, 21, 14, 13, 10, 18)
```

- Convert the strings into a factor

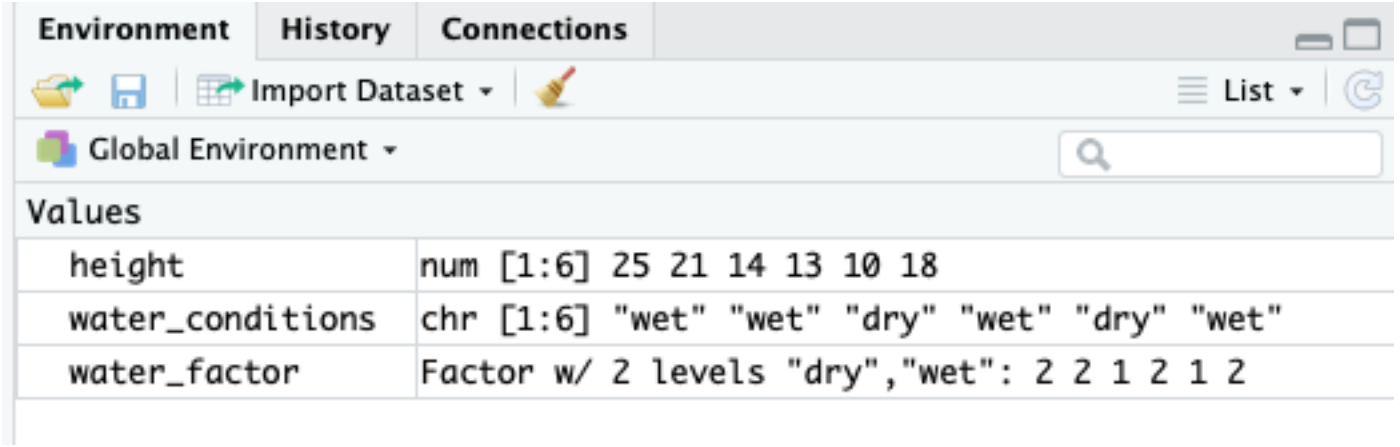
```
water_factor <- factor(water_conditions)
```

- Display the values of each data structure

```
water_conditions
water_factor
height
```

How to tell that a data structure is a factor

```
> water_conditions
[1] "wet" "wet" "dry" "wet" "dry" "wet"
> water_factor
[1] wet wet dry wet dry wet
Levels: dry wet
> height
[1] 25 21 14 13 10 18
>
```



- The main clue is that the **values of the levels** are listed.

Data frames and factors

- **character strings** imported from CSV files are automatically turned into **factors**
- **numbers** imported from CSV files are imported as **number vectors**
- This automatic behavior takes place because of the historical orientation of R towards statistics.
- The same behavior happens when data frames are built from individual vectors. (Investigate **`organism_info`** example from previous lesson)
- This can be good or bad depending on how you want to use the data.

Analysis of discontinuous data

Cockroach electroretinogram experiment

- See https://youtu.be/aAdnZsggZZw
- Difference in ability to detect colors of light

t-test exercise

Questions about the schools data

1. Is zip code a vector or a factor?
2. Should zip code be a vector or a factor?
3. Is school name a vector or a factor?
4. Should school name be a vector or a factor?

- Convert these data to the correct form using:

`factor()` turn a vector into a factor

`as.character()` turn a factor into a character vector

- How many levels of zip codes are there (vs. rows)?