**Cleaning data in R**

**Link :** <https://rstudio-pubs-static.s3.amazonaws.com/294962_40c4eb3b399d494b9f063e05847fa933.html>

**Chapter 1**

1. Collect data
2. Clean data
3. Analyze Data
4. Report Data

Data cleaning process

1. Exploring raw data
2. Tidying data
3. Preparing your data for analysis

Data that is missing for a reason 🡪 Missing not at Random (MNAR)

1. **Exploring raw data**

* Understand the structure of your data
* Check its class using **class(var)**
* View its dimensions using **dim(var)**
* Look at column names using **names(var)**
* Analyze structure using **str(var)**
* Load dplyr package using library(dplyr) and then view using **glimpse()**
* View a summary using **summary()**
* Look at your data
* View the top using **head()**
* View the bottom using **tail()**
* Visualize your data
* View histogram using **hist(var)**
* View plot of two variables (scatterplot) using **plot(x,y)**

1. **Tidy data**

Observations as rows, variables as columns, one unit per table.

Columns into key-value pairs

**gather(name of data, new name for column we are transposing, name of values, -column that were not touching OR that we are touching X1:X31)**

Key-value pairs into Columns

**spread(name of data, column you want to spread, values, -col)**

Separate columns

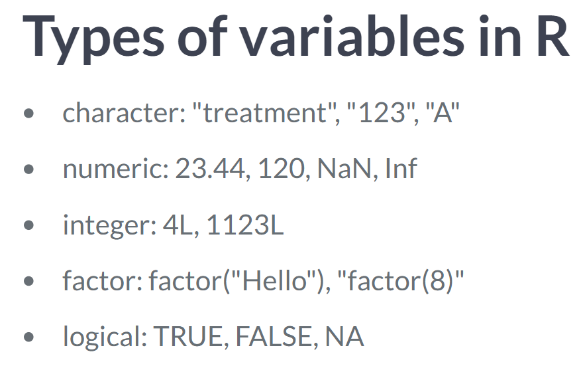
**separate(name of data, mixed column, into=c(“”,””), sep=”/”)**

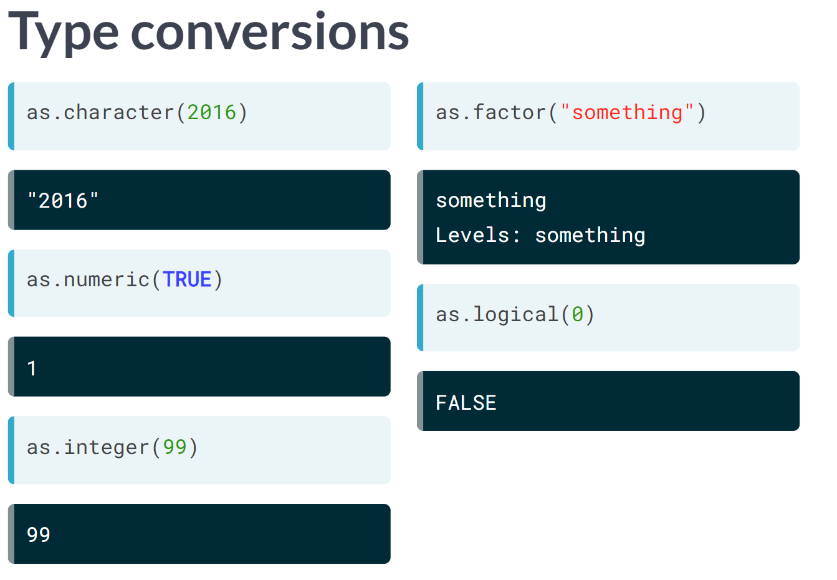
Unite columns

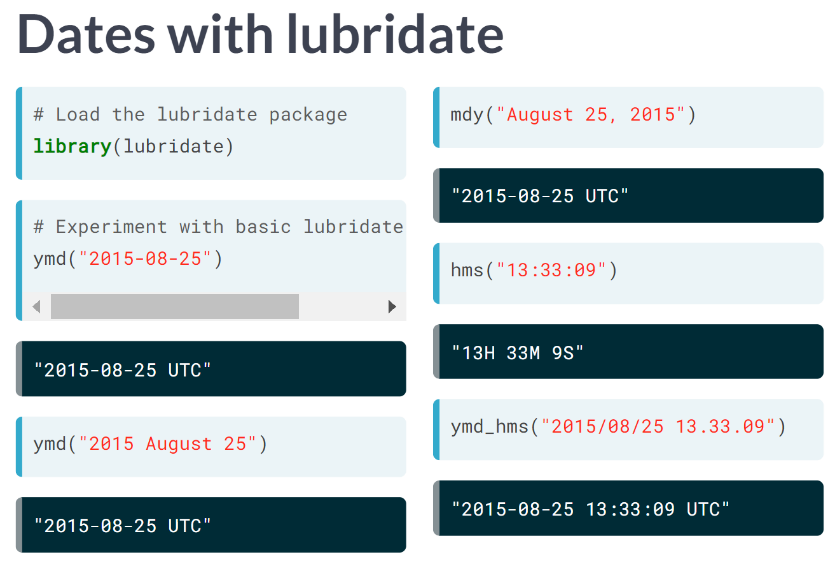
**unite(name of data, new name, old name, olda name, sep=”/”)**

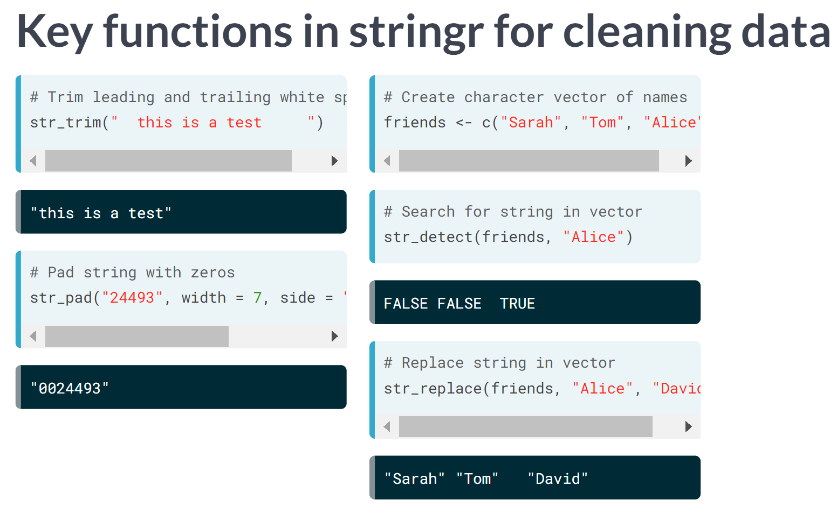
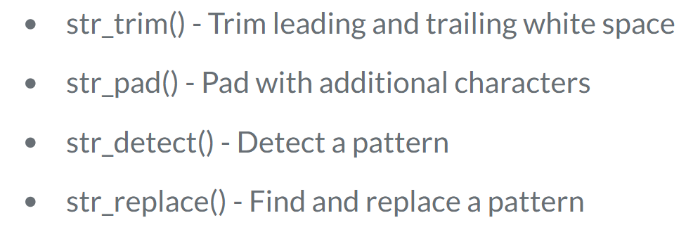
new name should be between “”

1. **Type conversions**



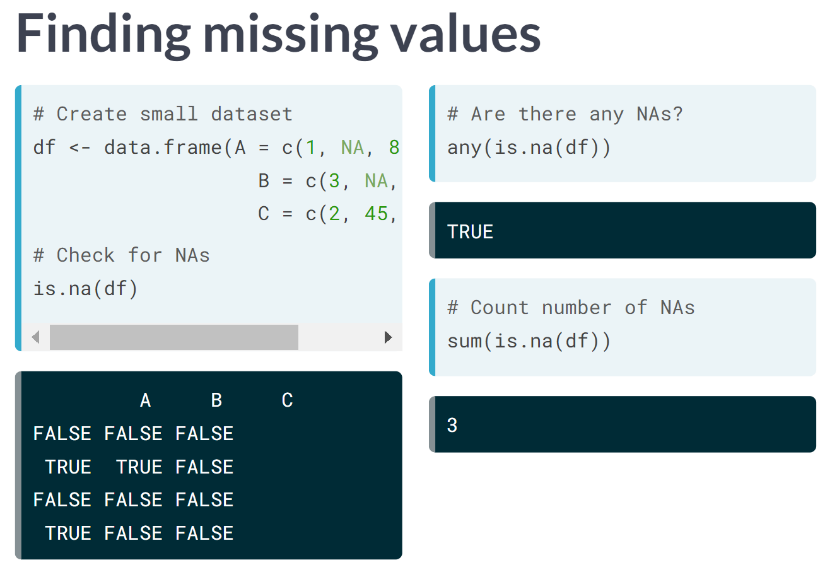
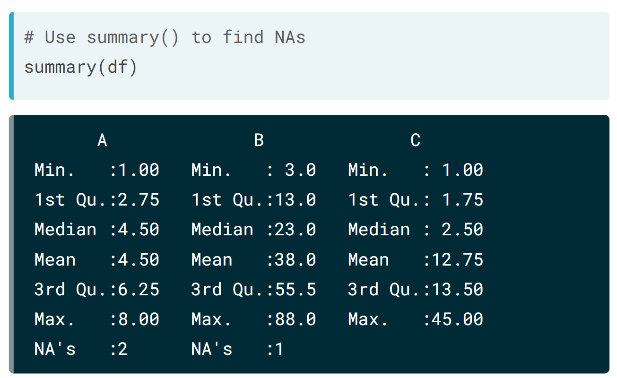
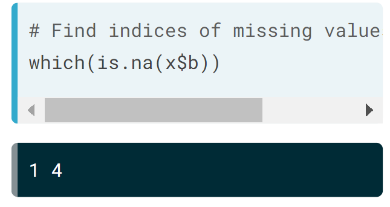


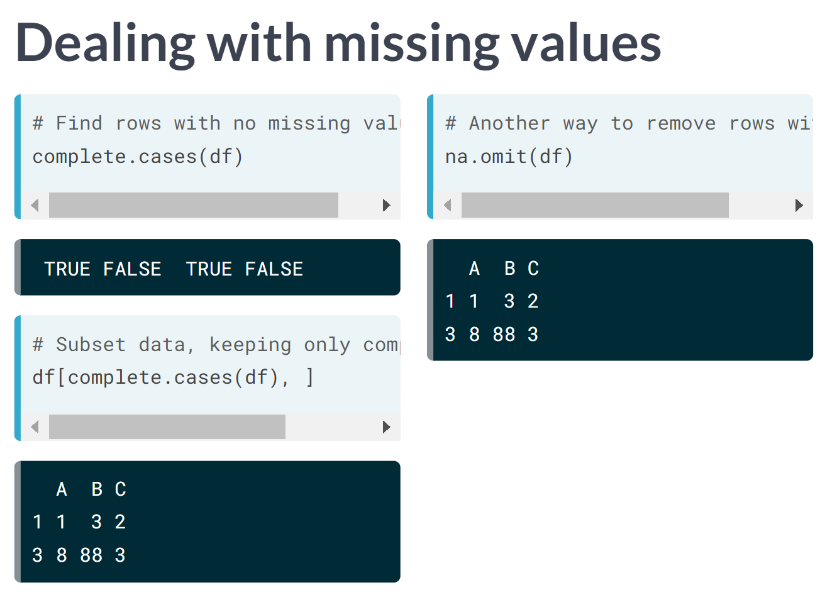




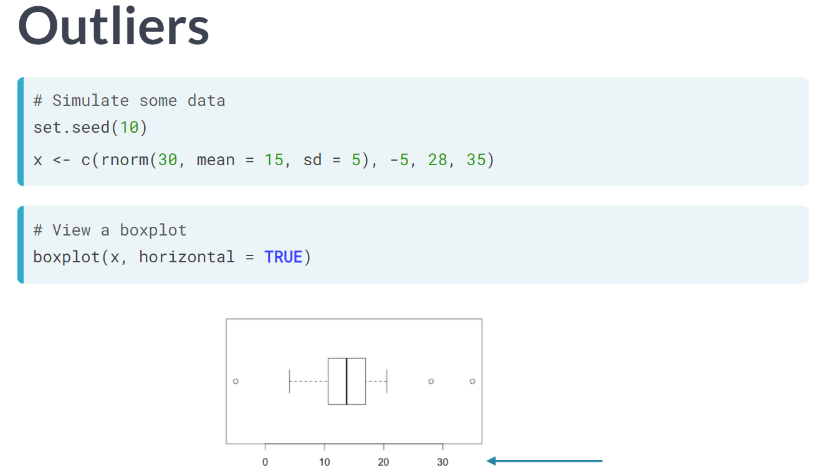
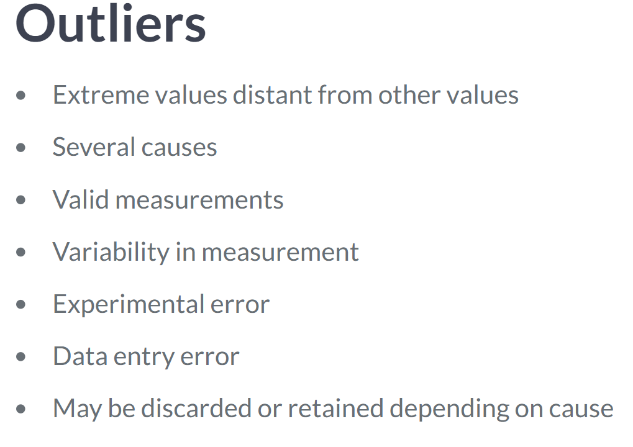
All to lower case 🡪 tolower()

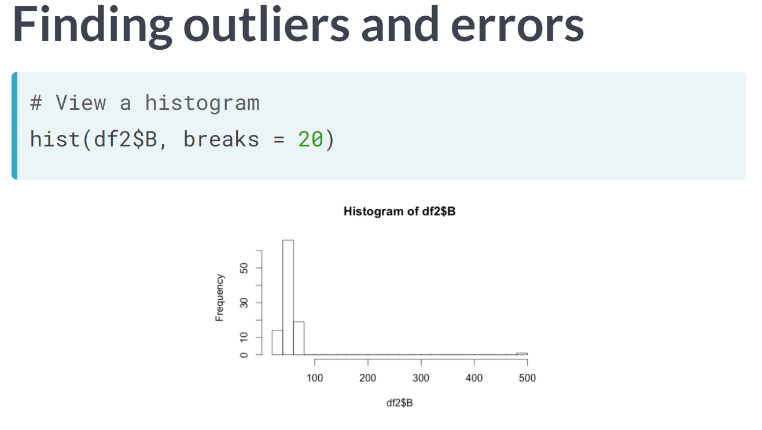
All to upper case 🡪 toupper()





# Replace all empty strings in status with NA using data$column[data$column =="word"] <- NA





Use summary() also

1. **Example**

# Verify that weather is a data.frame

class(weather)

# Check the dimensions

dim(weather)

# View the column names

names(weather)

# First remove column of row names

without\_x <- weather2[, -1]

# Convert characters to numerics

weather6 <- mutate\_at(weather5, vars(CloudCover:WindDirDegrees), funs(as.numeric))