Preparing Your Dataset for Analysis in R: A Quick Guide

Use this quick reference on the job to help recall functions and examples covered in Course 2: Data Manipulation and Cleaning in R.

Problem	Solution (Function)	Example
Missing values	is.na(), complete.cases()	is.na(df$price)
Removing missing data	na.omit()	df <- na.omit(df)
Extra whitespace	str_trim()	df$name <- str_trim(df$name)
Duplicate rows	distinct()	df <- distinct(df)
Incorrect data types	as.numeric(), as.character(), as.Date()	df$price <- as.numeric(df$price)
Extracting patterns from strings	str_extract(), str_extract_all()	df$zip <- str_extract(df$address, "\\d{5}")
Detecting patterns in strings	str_detect(), str_which()	df[str_detect(df$product, "TV"),]
Replacing/Substituting text in strings	str_replace(), str_replace_all()	df$phone <- str_replace_all(df$phone, "-", "")
Splitting strings	str_split(), str_split_fixed()	df$parts <- str_split_fixed(df$code, "-", 2)
Splitting combined fields	separate()	df <- separate(df, name, into = c("first", "last"), sep = " ")
Combining multiple columns	unite()	df <- unite(df, fullname, first, last, sep = " ")
Wide-to-long data shape	pivot_longer()	df_long <- pivot_longer(df, cols = Jan:Mar, names_to = "month", values_to = "sales")
Long-to-wide data shape	pivot_wider()	df_wide <- pivot_wider(df, names_from = month, values_from = sales)
Filtering data rows	filter()	filter(df, price > 100)
Selecting specific columns	select()	select(df, customer, order_date)
Creating new calculated columns	mutate()	df <- mutate(df, total = quantity * price)
Reordering data	arrange()	arrange(df, desc(order_date))
Outliers (extreme values) detection	min(), max(), summary()	summary(df$price)
Calculating variability	sd()	sd(df$price, na.rm = TRUE)
Summarizing groups	group_by(), summarize()	df %>% group_by(category) %>% summarize(avg_price = mean(price, na.rm=TRUE))