

# RとQuartoではじめるデータサイエンス《2026》

## #2 Rの基本的な操作方法(1)

荻谷千尋

Wed, 15, Apr, 2026

### 目次

1. 前回の振り返り
2. データフレームの構造と型
3. データの抽出
4. データの要約と集計
5. 生成AIを活用する
6. Rの構造：base R と modern R(tidyverse)の関係

### 0. 本日の目標

1. データフレームの基本的な構造（行・列・変数）を理解する
2. データの型（数値・文字・ファクターなど）の違いを**なんとなく**理解する
3. データの抽出・除外についての基本的な操作を行ってみる
4. データを簡単に計算し、その結果を図にしてみる（**詳しい説明、理解は次週以降**）
5. 生成AIを補助ツールとして活用する可能性と注意点を理解する
6. base Rとmodern Rの関係を**なんとなく**理解する

### 1. データフレームの構造と型

- データフレーム = データを表の形でまとめたもの
  - **明確な列名とそれに従った値からなる**
    - 行 (row) : 1つの観測 (例: 1羽のペンギン)
    - 列 (column) : 1つの変数 (例: 体重・種類)
- 型によって「できる操作」が異なる
  - 数値: 計算に向いている
  - 文字: 分類に向いている
  - カテゴリ: 分類に使われるデータ (順序を持つこともある)
- skim()関数によるデータの特徴・型の把握

### 2. データの抽出

- filter(): 指定した変数のデータのみ残す
- filter\_out(): 指定した変数のデータを除外する
- 欠損値
  - Rは**NAの扱いを明示しないと計算しない仕様**
  - 欠損値を除外するのが鉄則
  - 除外の仕方
    - is.na() → NAかどうかを調べる
    - ! → 否定 (~ではない)

- 例：`filter(!is.na(body_mass))`

### 3. データの要約と集計（基本編）

- 平均値を算出する：`summarise(avg_body_mass = mean(body_mass))`
- グループごとの平均値を算出する：`group_by(species)`

### 4. 生成AIを活用する

#### (1) 生成AIの利点を知る

- コードと生成AIは相性がよい

#### (2) 生成AIの限界を知る

- 古い情報・不正確な情報を出すことがある
- 存在しない関数、間違ったコードを出すことがある
- AIは目的や問い、仮説を立てることはできない
  - **[?] 何を分析したいかは自分で考える必要がある**

### 7. Rの構造：base R と modern R(tidyverse)の関係

#### (1) 3層構造

- Packages
  - 目的別に追加するツール：**家具や家電**
- modern R：tidyな（整然とした）R:
  - 近年の改良で、より読みやすく・便利になった文法：**家の骨組み・壁・柱、設計の考え**
- base R：R言語そのもの：**家の基礎部分**

#### (2) パイプ処理

- `|>` または `%>%`
- 処理を連続させるための記号（正確には、演算子という）
  - より正確には「前の結果を次の関数に渡して、処理を順番につなげるための記号」
  - **何をどうしたいのかが明確**（書きやすく、読みやすい）
  - **コードを単純化でき、上から順に読めるコードになる**

### 宿題

#### 授業の感想：

- 回答先：Google Forms
- 締め切り：4月17日（金）23時59分

#### 演習：

- 内容：授業内で課した演習①～④
- ファイル：演習の該当箇所を抜き出すのではなく、qmdファイルをそのまま提出
  - ファイル名：**氏名をかな表記**にして下さい（例：かりやちひろ.qmd）
- 回答先：dropbox
- 締め切り：4月20日（月）23時59分