RとQuartoではじめるデータサイエンス《2026》

#3 可視化(1)

苅谷千尋

Wed, 22, Apr, 2026

目次

0. 本日の目標

1. ggplotの基本構造（data；aes；geom）を**大まかに**理解する
2. 作図は「何を見たいのか」と「データの型」で決まることを理解する
3. 変数の数とデータの型に応じて、適切なgeom_関数を選びながら図を作る（サンプルコード参照）
4. 変数の数に応じて、可視化される情報量が異なることを**直感的に**理解する
5. レポートのためのデータセットに触れてみる（**宿題あり**）

2. ggplot

⑴ 3要素（3ページ「ggplot layer概念図」参照）

- data（材料）：
  1. 元データ（行データ）：1行が1観測（例：1羽、1人など）のデータ（例：penguins）
     - `geom_bar()`（件数カウント）；`geom_histogram()`（分布）；`geom_point()`（散布図）
  2. 集計済みデータ：平均・合計・割合などを計算した後のデータ
     - `geom_col()`
- aes（設計図）：
  - 美的マッピング（aes = aesthetic mapping）/ どのデータをどこに使うかというルール / 全てのレイヤーに共通して適用される
- geom（描く）：
  - 実際に図を描く部分 / 複数重ねることができる（棒＋点＋線など） / geom_○○という関数名をもつ
  - ジオメトリ・幾何学（geometry）に由来。「ジィーム」
  - `+` レイヤー演算子：グラフの要素を順番に重ねて追加する

⑵ 5 Named Graphs (5NG)

- 棒グラフ；ヒストグラム；箱ひげ図；散布図；折れ線グラフ

⑶ グラフの作り方：3つのポイント

1. グラフの種類ではなく、**どんな関係を見たいのか**を考える（例：分布；比較；関係；推移）
2. **データの型（数値（離散；連続）；カテゴリ（名義尺度；順序尺度））を把握する**
3. **元データのまま**可視化するのか；**集計してから**可視化するのか

⑷ 試行錯誤の二つの留意点

> **コードを使ってグラフを作るすばらしい点は、いったん壊したら元に戻せないような操作が含まれていないこと**です。もし何かがうまくいかなかったら、何が起きているかを特定し、それを修正し、そして作図コードをもう一度実行すればよいのです。
>
> 2つ目の点は、**ggplotを使った作業の主な流れはいつも同じ**だということです（ヒーリー, キーラン (2021), 124ページ）。

3. 単変量（1変数）の作図

- 1変数：カテゴリ：**棒グラフ**：`geom_bar()`
  - xごとに件数を集計して棒の高さにする（自動カウント・**元データを使用する**）
- 1変数：連続値：**ヒストグラム**：`geom_histogram()`
  - 数値の「分布」を見る / 値を区間（ビン）に分けて数える

4. 二変量（2変数）の作図

- 2変数：数値 × 数値：**散布図**：`geom_point()`
  - 2つの数値変数の関係（傾向・ばらつき）を可視化
- 2変数：カテゴリ×カテゴリ：**積み上げ棒グラフ**：`geom_bar(position = "stack")`（デフォルト)
- 2変数：カテゴリ×カテゴリ：**並列棒グラフ**：`geom_bar(position = "dodge")`
  - dodge（避ける、かわすの意）：重なりを避けて横に配置
- 2変数：カテゴリ×カテゴリ：**帯グラフ**：`geom_bar(position = "fill")`
  - 各カテゴリの構成比（割合）で表示
- 2変数：連続値×カテゴリ：**ヒストグラム**
  - ggplotに渡す前にグループ分け（`group_by`）を行い、`facet_wrap`で分ける
- 2変数：カテゴリ×数値：**棒グラフ** `geom_col()`
  - カテゴリ別に集計した数値を可視化
  - dplyr系の関数（`summarise()`など）についての知識を要する
- 2変数：連続値（または順序） × 数値：**折れ線グラフ** `geom_line()`

5. 多変量（3変数）の作図

- 複数の変数を加えることで情報量が増える
  - → **関係や違い、各変数の特徴が見えやすくなり、新たな疑問や仮説が生まれる**
- ggplotによる多変量の可視化
  - 色で区別する：colour（点・線）、fill（面）；図を分ける：facet

6. レポートのためのデータセット探し

- SSDSE（教育用標準データセット）
  - 「主要な公的統計を地域別に一覧できる表形式のデータセット」
  - **1行目を除外すれば、そのままRで使える**
  - **このデータセットからレポートを作成することを推奨**（他を使いたい人は要相談）

宿題

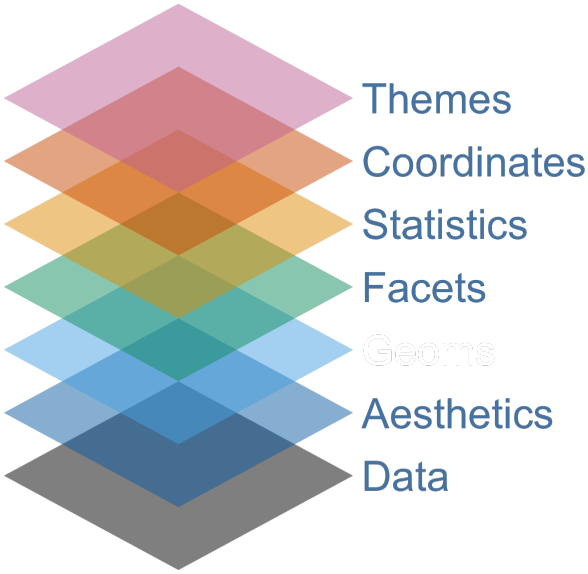
授業の感想：

- 授業の感想：
  - 回答先：Google Forms
  - 締め切り：4月24日（金）23時59分
- レポートに使いたいデータセット
  - 回答先：Google Forms
  - 締め切り：4月30日（木）10時30分

演習：

- 内容：演習④⑥
- 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：どこかに*__氏名を特定できる文字__を入れて下さい（例：inClassExercise_**kariya**.qmd）
- 回答先：dropbox
- 締め切り：4月27日（月）23時59分

ggplot layer概念図



- SOURCE: Data Science with R, Grammar of graphics

Reference

ヒーリー，キーラン (2021)『データ分析のためのデータ可視化入門』, 講談社.