Transforming Data in R

EES 4891/5891

Probability & Statistics for Geosciences

Jonathan Gilligan

Class #7: Tuesday, January 28 2025

Learning Goals

Learning Goals

1. Basics of writing code in R
   a. Naming variables and assigning values
   b. Calling functions
2. Working with Data in R
   a. Know what data frames are and how they're structures
   b. Know how to manipulate the rows of a data frame
      - Sorting (`arrange`)
      - Selecting (`filter`, `distinct`, `slice_`)
      - Summarizing (`summarize`, `count`)
   c. Know how to manipulate the columns of a data frame
      - Creating or changing columns (`mutate`)
      - Choosing, renaming, and rearrranging: (`select`, `rename`, `relocate`)
        - Auxiliary functions (`any_of`, `starts_with`, `ends_with`, ...)
   d. Know how to use pipes to combine transformations (`|>`)
   e. Know how to apply transformations to groups of rows (`group_by`, `ungroup`, `.by`)

R Workflow

R Workflow

- Scripts vs. Notebooks
  - Scripts: R code that you can run on its own
    - `.R` or `.r` files
  - Notebooks: R code integrated with text that you *render* to produce a document (web page, report, book, presentation, etc.)
    - `.Rmd` or `.qmd` files
    - Code blocks look like

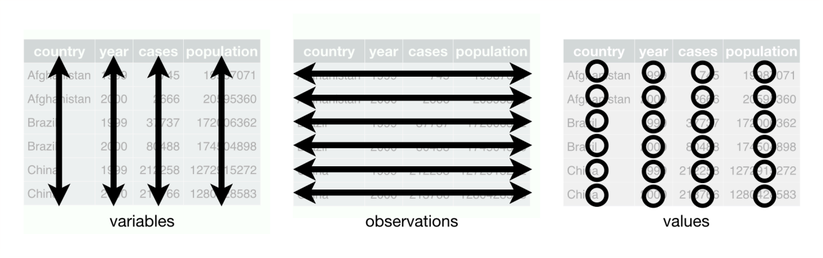
````
```{r block_name}
# R code goes here.
```
````

- Parts of an R script or code block:
  - Variables: name starts with `.`, `_`, or a letter, followed by any combination of letters, digits, `.`, or `_`.
    - e.g., `sample_12.3_weight` or `s_12.3_wt`
    - Assign values with `=` or `<-`
  - Functions:
    - `seq(0, 10, 2)`
    - `seq(from = 0, to = 10, by = 2)`

Transforming Data

Transforming Data

- `data.frame` and `tibble` objects



- Example

```
library(tidyverse)
library(fossil)
data(fdata.list)
fdata <- as_tibble(fdata.list)
```

  - You may need to run `install.packages("fossil")` in the RStudio console.

```
head(fdata)
```

```
## # A tibble: 6 × 5
##   locality species abundance longitude latitude
##   <fct>    <fct>        <int>     <dbl>    <dbl>
## 1 locA     spp1             1      -109       47
## 2 locA     spp8             1      -109       47
## 3 locA     spp12            1      -109       47
## 4 locB     spp5             1       -90       45
## 5 locB     spp6             5       -90       45
## 6 locB     spp8             5       -90       45
```

```
glimpse(fdata)
```

```
## Rows: 52
## Columns: 5
## $ locality  <fct> locA, locA, locA, locB, locB, …
## $ species   <fct> spp1, spp8, spp12, spp5, spp6,…
## $ abundance <int> 1, 1, 1, 1, 5, 5, 1, 4, 1, 1, …
## $ longitude <dbl> -109, -109, -109, -90, -90, -9…
## $ latitude  <dbl> 47, 47, 47, 45, 45, 45, 45, 45…
```

Jura Data Set

Jura Data Set

- Survey of soil contamination in the Swiss Jura
  - O. Attela, J.-P. Dubois, & R. Webster. 1994. *Environ. Pollution* **86**, 315.

```
library(tidyverse)
library(gstat)
data(jura)
jura <- as_tibble(jura.val) |>
  select(-(Xloc:Yloc))
```

- You may need to run `install.packages("gstat")` in the RStudio console.

```
glimpse(jura)
```

```
## Rows: 100
## Columns: 13
## $ Xloc    <dbl> 2.672, 3.589, 4.010, 2.942, 1.40…
## $ Yloc    <dbl> 3.558, 4.443, 4.713, 3.137, 2.74…
## $ long    <dbl> 6.854080, 6.865951, 6.871425, 6.…
## $ lat     <dbl> 47.14342, 47.15144, 47.15390, 47…
## $ Landuse <fct> Meadow, Meadow, Pasture, Pasture…
## $ Rock    <fct> Quaternary, Argovian, Argovian, …
## $ Cd      <dbl> 1.570, 2.045, 1.203, 0.490, 0.69…
## $ Co      <dbl> 8.28, 10.80, 12.00, 10.92, 8.12,…
## $ Cr      <dbl> 37.12, 40.80, 53.20, 23.40, 27.1…
## $ Cu      <dbl> 18.600, 11.480, 13.040, 5.640, 1…
## $ Ni      <dbl> 18.60, 21.52, 23.92, 14.60, 14.6…
## $ Pb      <dbl> 38.20, 33.36, 26.56, 25.88, 31.1…
## $ Zn      <dbl> 65.20, 112.80, 91.60, 41.20, 50.…
```

Transforming Rows

Transforming Rows

- Selecting:

```
  filter(jura, Landuse == "Meadow") |> head()
```

```
## # A tibble: 6 × 11
##    long   lat Landuse Rock            Cd    Co    Cr    Cu    Ni    Pb
Zn
##   <dbl> <dbl> <fct>   <fct>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl>
## 1  6.85  47.1 Meadow  Quaternary   1.57   8.28  37.1  18.6  18.6  38.2
65.2
## 2  6.87  47.2 Meadow  Argovian     2.04  10.8   40.8  11.5  21.5  33.4
113.
## 3  6.84  47.1 Meadow  Sequanian    0.692  8.12  27.2  10.3  14.6  31.2
50.4
## 4  6.85  47.1 Meadow  Kimmeridgian 0.92  10.6   49.0  30.3  31.5  68.1
103.
## 5  6.85  47.1 Meadow  Argovian     0.495  8.52  31.4  17.1  16.1  46.8
57.6
## 6  6.84  47.1 Meadow  Sequanian    1.19   9.68  37.4  31.4  22.4  72.4
108.
```

- Sorting:

```
filter(jura, Landuse == "Meadow") |>
  arrange(Rock, long, lat) |>
  head()
```

```
## # A tibble: 6 × 11
##    long   lat Landuse Rock        Cd    Co    Cr    Cu    Ni    Pb    Zn
##   <dbl> <dbl> <fct>   <fct>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  6.83  47.1 Meadow  Argovian 0.475  3.96  22.2  22.7  7.92  55.2  46.8
## 2  6.84  47.1 Meadow  Argovian 3.78   9.68  42.8  32.8 23.5   94.4 175.
## 3  6.84  47.1 Meadow  Argovian 0.585  5.8   39.9  15.2 13.2   56.4  51.2
## 4  6.84  47.1 Meadow  Argovian 0.57   4.08  24.9  21.4  9.68  67.2  56.8
```

Transforming Rows

- Selecting on multiple criteria:

```
  filter(jura, Landuse == "Meadow", Rock == "Quaternary", Cd >
         1.3) |>
    head()
```

```
## # A tibble: 5 × 11
##    long   lat Landuse Rock          Cd    Co    Cr    Cu    Ni    Pb
Zn
##   <dbl> <dbl> <fct>   <fct>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl>
## 1  6.85  47.1 Meadow  Quaternary  1.57  8.28  37.1  18.6  18.6  38.2
65.2
## 2  6.86  47.2 Meadow  Quaternary  1.58  5.8   40.4  56.4  22.5  93.6 109.
## 3  6.88  47.1 Meadow  Quaternary  1.42 11.1   27.5  18.8  20.6  36.5
63.2
## 4  6.86  47.1 Meadow  Quaternary  2.08 13.2   45.9  39    26.4  52.4 104
## 5  6.88  47.1 Meadow  Quaternary  2.61 20.6   37.2  24    29.4  47.2
86.4
```

Transforming Columns

Transforming Columns

```
mutate(jura, CuNi = Cu / Ni, PbZn = Pb/Zn) |>
rename(longitude = long, latitude = lat) |>
  head()
```

```
## # A tibble: 6 × 13
##   longitude latitude Landuse Rock           Cd    Co    Cr    Cu    Ni    Pb    Zn  CuNi
PbZn
##       <dbl>    <dbl> <fct>   <fct>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl>
## 1      6.85     47.1 Meadow  Quaternary   1.57   8.28  37.1 18.6   18.6  38.2  65.2 1
0.586
## 2      6.87     47.2 Meadow  Argovian     2.04  10.8   40.8 11.5   21.5  33.4 113.  0.533
0.296
## 3      6.87     47.2 Pasture Argovian     1.20  12     53.2 13.0   23.9  26.6  91.6 0.545
0.290
## 4      6.86     47.1 Pasture Quaternary   0.49  10.9   23.4  5.64  14.6  25.9  41.2 0.386
0.628
## 5      6.84     47.1 Meadow  Sequanian    0.692  8.12  27.2 10.3   14.6  31.2  50.4 0.705
0.618
## 6      6.87     47.1 Forest  Kimmeridgian 1.75   9.12  35.5  8.36  26.4  37.7  63.2 0.317
0.597
```

```
mutate(jura, CuNi = Cu / Ni, PbZn = Pb/Zn) |>
rename(longitude = long, latitude = lat) |>
relocate(CuNi:PbZn, .before = Cd) |>
  head()
```

```
## # A tibble: 6 × 13
##   longitude latitude Landuse Rock          CuNi  PbZn    Cd    Co    Cr    Cu    Ni    Pb
Zn
##       <dbl>    <dbl> <fct>   <fct>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
<dbl>
## 1       6.85     47.1 Meadow  Quaternary   1     0.586 1.57   8.28  37.1 18.6   18.6  38.2
65.2
## 2       6.87     47.2 Meadow  Argovian     0.533 0.296 2.04  10.8   40.8 11.5   21.5  33.4
113.
## 3       6.87     47.2 Pasture Argovian     0.545 0.290 1.20  12     53.2 13.0   23.9  26.6
91.6
## 4       6.86     47.1 Pasture Quaternary   0.386 0.628 0.49  10.9   23.4  5.64  14.6  25.9
41.2
## 5       6.84     47.1 Meadow  Sequanian    0.705 0.618 0.692  8.12  27.2 10.3   14.6  31.2
50.4
## 6       6.87     47.1 Forest  Kimmeridgian 0.317 0.597 1.75   9.12  35.5  8.36  26.4  37.7
63.2
```

Pipes

Pipes

- The pipe `|>` sends the output from one function into another.
- This code is hard to read:

```
relocate( rename( mutate( jura,
                          CuNi = Cu / Ni,
                          PbZn = Pb/Zn),
                  longitude = long,
                  latitude = lat),
          CuNi:PbZn, .before = Cd)
```

- This code is much easier to read

```
jura |>
  mutate(CuNi = Cu / Ni, PbZn = Pb/Zn) |>
  rename(longitude = long, latitude =
         lat) |>
  relocate(CuNi:PbZn, .before = Cd)
```

- We could also write

```
tmp <- mutate(jura, CuNi = Cu / Ni, PbZn =
         Pb/Zn)
tmp <- rename(tmp, longitude = long,
              latitude = lat)
relocate(tmp, CuNi:PbZn, .before = Cd)
```

Summarizing and Grouping

Summarizing and Grouping

More Grouping

- Average lead content:

```
jura |> summarize(mean_Pb = mean(Pb),
                  sd_Pb = sd(Pb),
                  count = n())
```

```
## # A tibble: 1 × 3
##   mean_Pb sd_Pb count
##     <dbl> <dbl> <int>
## 1    56.5  40.5   100
```

- Grouped Summaries:

```
jura |> group_by(Landuse, Rock) |>
  summarize(mean_Pb = mean(Pb),
            sd_Pb = sd(Pb), count =
          n()) |>
  ungroup()
```

```
## # A tibble: 15 × 5
##    Landuse Rock          mean_Pb sd_Pb count
##    <fct>   <fct>           <dbl> <dbl> <int>
##  1 Forest  Argovian         33.9  3.21     4
##  2 Forest  Kimmeridgian     53.6 15.3     10
##  3 Forest  Sequanian        57.9 13.4      2
##  4 Forest  Portlandian      42.0  9.56     2
##  5 Pasture Argovian         28.5 11.4      4
##  6 Pasture Kimmeridgian     60.9 69.4     14
##  7 Pasture Sequanian        42.1 11.5      6
##  8 Pasture Quaternary       37.1 15.9      2
##  9 Meadow  Argovian         53.1 25.3     14
## 10 Meadow  Kimmeridgian     50.1 11.9     13
## 11 Meadow  Sequanian        73.1 54.1     18
## 12 Meadow  Portlandian     109.  NA         1
## 13 Meadow  Quaternary       60.8 41.5      7
## 14 Tillage Argovian        153.  NA         1
## 15 Tillage Kimmeridgian     42.3  5.77     2
```

Slicing

- Counting combinations:

```
jura |> count(Landuse, Rock)
```

```
## # A tibble: 15 × 3
##    Landuse Rock            n
##    <fct>   <fct>       <int>
##  1 Forest  Argovian        4
##  2 Forest  Kimmeridgian   10
##  3 Forest  Sequanian       2
##  4 Forest  Portlandian     2
##  5 Pasture Argovian        4
##  6 Pasture Kimmeridgian   14
##  7 Pasture Sequanian       6
##  8 Pasture Quaternary      2
##  9 Meadow  Argovian       14
## 10 Meadow  Kimmeridgian   13
## 11 Meadow  Sequanian       18
## 12 Meadow  Portlandian     1
## 13 Meadow  Quaternary      7
## 14 Tillage Argovian       1
## 15 Tillage Kimmeridgian    2
```

- Alternate Grouping:

```
jura |> summarize(mean_Pb = mean(Pb),
                  sd_Pb = sd(Pb),
                  count = n(),
                  .by = c("Landuse",
        "Rock"))
```

```
## # A tibble: 15 × 5
##    Landuse Rock          mean_Pb sd_Pb count
##    <fct>   <fct>           <dbl> <dbl> <int>
##  1 Meadow  Quaternary       60.8 41.5      7
##  2 Meadow  Argovian         53.1 25.3     14
##  3 Pasture Argovian         28.5 11.4      4
##  4 Pasture Quaternary       37.1 15.9      2
##  5 Meadow  Sequanian        73.1 54.1     18
##  6 Forest  Kimmeridgian     53.6 15.3     10
##  7 Pasture Sequanian        42.1 11.5      6
##  8 Meadow  Kimmeridgian     50.1 11.9     13
##  9 Pasture Kimmeridgian     60.9 69.4     14
## 10 Forest  Portlandian      42.0  9.56     2
## 11 Forest  Argovian         33.9  3.21     4
## 12 Meadow  Portlandian     109.  NA         1
## 13 Tillage Kimmeridgian     42.3  5.77     2
## 14 Forest  Sequanian        57.9 13.4      2
## 15 Tillage Argovian        153.  NA         1
```

- Selecting:

```
  jura |> group_by(Landuse, Rock) |>
    slice_max(Pb, n = 1)
```

```
## # A tibble: 15 × 11
## # Groups:   Landuse, Rock [15]
##     long   lat Landuse Rock            Cd    Co    Cr     Cu    Ni
Pb    Zn
##    <dbl> <dbl> <fct>   <fct>        <dbl> <dbl> <dbl>  <dbl> <dbl>
<dbl> <dbl>
##  1  6.86  47.2 Forest  Argovian     1.32   3.74  27.6   5.4  14.4
37.5  46.4
##  2  6.86  47.1 Forest  Kimmeridgian 1.25   8.08  39.6  13.1  18.6   88
86.8
##  3  6.85  47.1 Forest  Sequanian    1.01   9.96  28.7   5.96 17.4
67.4  52.5
##  4  6.87  47.1 Forest  Portlandian  1.22   5.24  27.0   5.52 21.0
48.8  46.4
##  5  6.85  47.1 Pasture Argovian     0.375 12.0   34.1  19.4  16.4
45.2  70
##  6  6.86  47.1 Pasture Kimmeridgian 1.76  10.3   40.5 127    30.8  300
192
##  7  6.83  47.1 Pasture Sequanian    2.54  12.6   70     8.72 26.2
55.6  71.6
##  8  6.88  47.1 Pasture Quaternary   1.31  12.7   34.8  17.7  19.6
48.4  80.4
##  9  6.85  47.1 Meadow  Argovian     0.394  4.44  21.6  39.6   8.92 106.
72.4
## 10  6.85  47.1 Meadow  Kimmeridgian 0.825 15.3   36.5  31.2  25.4
70.4  75.4
## 11  6.86  47.1 Meadow  Sequanian    1.78  11.4   41   155.   24.5  240.
260.
## 12  6.88  47.1 Meadow  Portlandian  1.62  12.0   34.6  91.2  30.2  109.
157.
## 13  6.85  47.1 Meadow  Quaternary   0.75  15.6   29.8  73.1  20.2  139.
95.7
## 14  6.85  47.1 Tillage Argovian     1.31   8.44  41.6 118.   20.4  153.
145.
## 15  6.87  47.1 Tillage Kimmeridgian 1.93  13.8   45    19.3  35.7
46.4  90
```

- Random Sampling

```
  jura |> slice_sample(n = 5)
```

```
## # A tibble: 5 × 11
##    long   lat Landuse Rock            Cd    Co    Cr    Cu    Ni    Pb
Zn
##   <dbl> <dbl> <fct>   <fct>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl>
## 1  6.87  47.1 Meadow  Kimmeridgian 0.855 12.4  32.9  25.5  22.3   51.6
80.4
## 2  6.85  47.1 Meadow  Argovian     0.495  8.52 31.4  17.1  16.1   46.8
57.6
## 3  6.88  47.1 Meadow  Quaternary   2.61  20.6  37.2  24    29.4   47.2
86.4
## 4  6.87  47.1 Forest  Kimmeridgian 0.51   1.65  3.32  5.96  1.98  62.4
60.4
## 5  6.85  47.1 Pasture Sequanian    1.66  14.8  40.3  37    30.1   43.8
87.8
```

Discuss Homework