Lecture 12

2024-10-14

Tidying and Joining Data

Pivot Longer

First let’s load our packages:

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(ps270data)
mortality
```

```
## # A tibble: 217 x 52
##    country      country_code indicator ‘1972‘ ‘1973‘ ‘1974‘ ‘1975‘ ‘1976‘ ‘1977‘
##    <chr>        <chr>        <chr>      <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Aruba        ABW          Mortalit~   NA     NA     NA     NA     NA     NA
##  2 Afghanistan  AFG          Mortalit~  291    285.   280.   274.   268    262.
##  3 Angola       AGO          Mortalit~   NA     NA     NA     NA     NA     NA
##  4 Albania      ALB          Mortalit~   NA     NA     NA     NA     NA     NA
##  5 Andorra      AND          Mortalit~   NA     NA     NA     NA     NA     NA
##  6 United Arab~ ARE          Mortalit~   80.1   72.6   65.7   59.4   53.6   48.3
##  7 Argentina    ARG          Mortalit~   69.7   68.2   66.1   63.3   59.8   55.7
##  8 Armenia      ARM          Mortalit~   NA     NA     NA     NA     87.1   83.6
##  9 American Sa~ ASM          Mortalit~   NA     NA     NA     NA     NA     NA
## 10 Antigua and~ ATG          Mortalit~   26.9   25.1   23.5   22.1   20.8   19.5
## # i 207 more rows
## # i 43 more variables: ‘1978‘ <dbl>, ‘1979‘ <dbl>, ‘1980‘ <dbl>, ‘1981‘ <dbl>,
## #   ‘1982‘ <dbl>, ‘1983‘ <dbl>, ‘1984‘ <dbl>, ‘1985‘ <dbl>, ‘1986‘ <dbl>,
## #   ‘1987‘ <dbl>, ‘1988‘ <dbl>, ‘1989‘ <dbl>, ‘1990‘ <dbl>, ‘1991‘ <dbl>,
## #   ‘1992‘ <dbl>, ‘1993‘ <dbl>, ‘1994‘ <dbl>, ‘1995‘ <dbl>, ‘1996‘ <dbl>,
## #   ‘1997‘ <dbl>, ‘1998‘ <dbl>, ‘1999‘ <dbl>, ‘2000‘ <dbl>, ‘2001‘ <dbl>,
## #   ‘2002‘ <dbl>, ‘2003‘ <dbl>, ‘2004‘ <dbl>, ‘2005‘ <dbl>, ‘2006‘ <dbl>, ...
```

to convert a data set into the long format, use the pivot_longer() function

mydata |> pivot_longer(cols = , names_to = , values_to =)

Let's do it with the mortality data

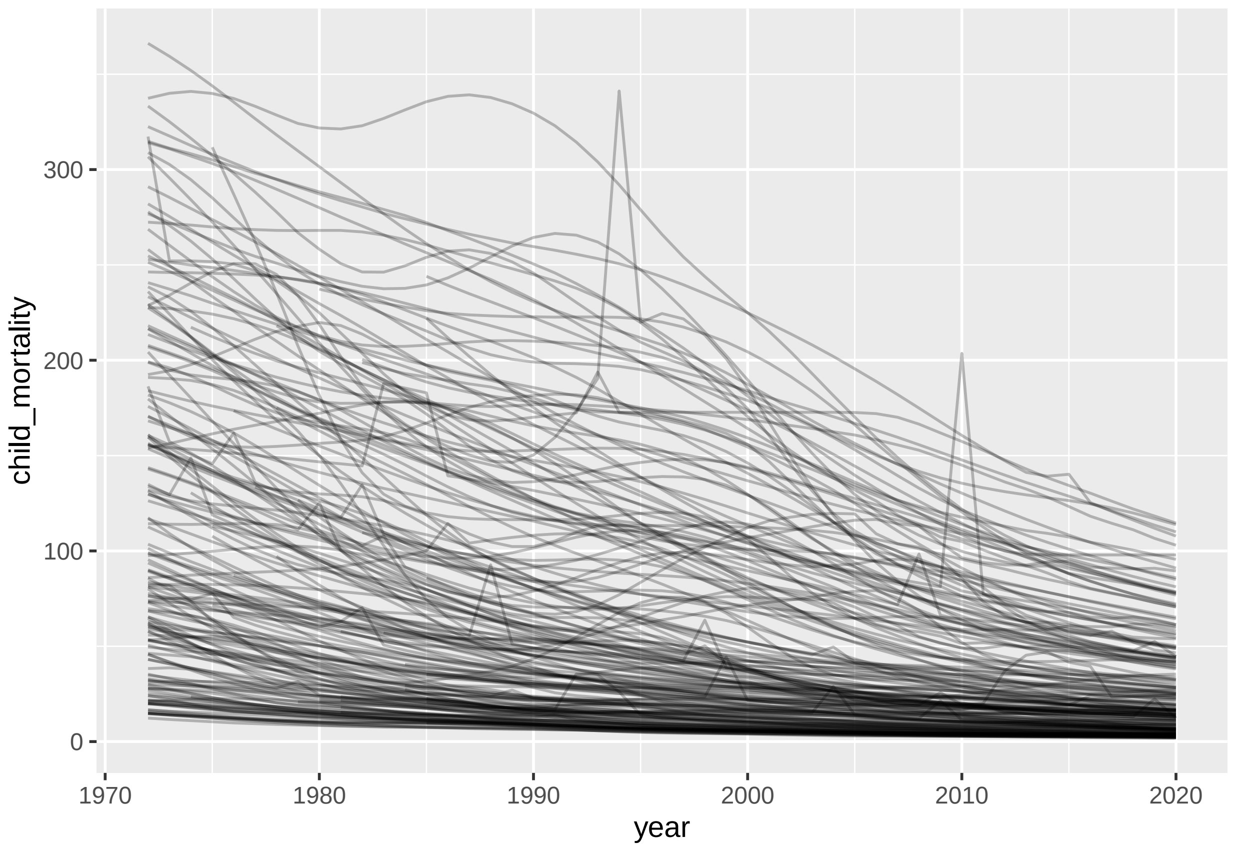
```
mortality |>
  select(-indicator) |>
  pivot_longer(
    cols = `1972`:`2020`,
    names_to = "year",
    values_to = "child_mortality"
  )
```

```
## # A tibble: 10,633 x 4
##    country country_code year  child_mortality
##    <chr>   <chr>        <chr>           <dbl>
##  1 Aruba   ABW          1972               NA
##  2 Aruba   ABW          1973               NA
##  3 Aruba   ABW          1974               NA
##  4 Aruba   ABW          1975               NA
##  5 Aruba   ABW          1976               NA
##  6 Aruba   ABW          1977               NA
##  7 Aruba   ABW          1978               NA
##  8 Aruba   ABW          1979               NA
##  9 Aruba   ABW          1980               NA
## 10 Aruba   ABW          1981               NA
## # i 10,623 more rows
```

let's do a line plot

```
mortality |>
  select(-indicator) |>
  pivot_longer(
    cols = `1972`:`2020`,
    names_to = "year",
    values_to = "child_mortality"
  )  |>
  mutate(year = as.integer(year)) |>
  ggplot(mapping = aes(x = year, y = child_mortality, group = country)) +
  geom_line(alpha = 0.25)
```

```
## Warning: Removed 1476 rows containing missing values or values outside the scale range
## (‘geom_line()‘).
```



let's practice pivot_longer on another dataset

```
spotify
```

```
## # A tibble: 490 x 54
##    `Track Name`     Artist week1 week2 week3 week4 week5 week6 week7 week8 week9
##    <chr>            <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 The Box          Roddy~     1     1     1     1     1     1     1     1     1
##  2 ROXANNE          Arizo~     2     4     5     4     4     4     6     7     9
##  3 Yummy            Justi~     3     6    17    17    17    24    15    32    NA
##  4 Circles          Post ~     4     7     9    10     7    10    11    10    17
##  5 BOP              DaBaby     5     5     7     5    11    12    18    18    32
##  6 Falling          Trevo~     6     8    10     7     6     8    10    11    18
##  7 Dance Monkey     Tones~     7    13    13    12    12    13    17    13    21
##  8 Bandit (with Yo~ Juice~     8    11    14    14    15    20    27    26    42
##  9 Futsal Shuffle ~ Lil U~     9     9    19    21    24    32    40    49    NA
## 10 everything i wa~ Billi~    10    17    28     9     8    11    14    17    29
## # i 480 more rows
## # i 43 more variables: week10 <dbl>, week11 <dbl>, week12 <dbl>, week13 <dbl>,
## #   week14 <dbl>, week15 <dbl>, week16 <dbl>, week17 <dbl>, week18 <dbl>,
## #   week19 <dbl>, week20 <dbl>, week21 <dbl>, week22 <dbl>, week23 <dbl>,
## #   week24 <dbl>, week25 <dbl>, week26 <dbl>, week27 <dbl>, week28 <dbl>,
## #   week29 <dbl>, week30 <dbl>, week31 <dbl>, week32 <dbl>, week33 <dbl>,
## #   week34 <dbl>, week35 <dbl>, week36 <dbl>, week37 <dbl>, week38 <dbl>, ...
```

```r
spotify |>
  pivot_longer(cols = c(-`Track Name`, -`Artist`),
               names_to = "week_of_year",
               values_to = "rank",
               names_prefix = "week") |>
  mutate(week_of_year = as.integer(week_of_year))
```

```
## # A tibble: 25,480 x 4
##    `Track Name` Artist      week_of_year  rank
##    <chr>        <chr>              <int> <dbl>
##  1 The Box      Roddy Ricch            1     1
##  2 The Box      Roddy Ricch            2     1
##  3 The Box      Roddy Ricch            3     1
##  4 The Box      Roddy Ricch            4     1
##  5 The Box      Roddy Ricch            5     1
##  6 The Box      Roddy Ricch            6     1
##  7 The Box      Roddy Ricch            7     1
##  8 The Box      Roddy Ricch            8     1
##  9 The Box      Roddy Ricch            9     1
## 10 The Box      Roddy Ricch           10     1
## # i 25,470 more rows
```

Joining Data Sets

```r
library(gapminder)
gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # i 1,694 more rows
```

first, assign our pivoted mortality data to the object mortality_long

```r
mortality_long <- mortality |>
  select(-indicator) |>
  pivot_longer(
    cols = `1972`:`2020`,
    names_to = "year",
    values_to = "child_mortality"
  ) |>
  mutate(year = as.integer(year))
```

```
mortality_long
```

```
## # A tibble: 10,633 x 4
##    country country_code  year child_mortality
##    <chr>   <chr>        <int>           <dbl>
##  1 Aruba   ABW           1972              NA
##  2 Aruba   ABW           1973              NA
##  3 Aruba   ABW           1974              NA
##  4 Aruba   ABW           1975              NA
##  5 Aruba   ABW           1976              NA
##  6 Aruba   ABW           1977              NA
##  7 Aruba   ABW           1978              NA
##  8 Aruba   ABW           1979              NA
##  9 Aruba   ABW           1980              NA
## 10 Aruba   ABW           1981              NA
## # i 10,623 more rows
```

Check that keys are unique

```
gapminder |>
  count(country, year) |>
  filter(n > 1)
```

```
## # A tibble: 0 x 3
## # i 3 variables: country <fct>, year <int>, n <int>
```

same for the other data

```
mortality_long |>
  count(country, year) |>
  filter(n > 1)
```

```
## # A tibble: 0 x 3
## # i 3 variables: country <chr>, year <int>, n <int>
```

first we use the left_join() function

```
gapminder |>
  left_join(mortality_long)
```

```
## Joining with `by = join_by(country, year)`
```

```
## # A tibble: 1,704 x 8
##    country continent  year lifeExp    pop gdpPercap country_code child_mortality
##    <chr>   <fct>     <int>   <dbl>  <int>     <dbl> <chr>                  <dbl>
##  1 Afghan~ Asia       1952    28.8 8.43e6      779.  <NA>                     NA
##  2 Afghan~ Asia       1957    30.3 9.24e6      821.  <NA>                     NA
##  3 Afghan~ Asia       1962    32.0 1.03e7      853. <NA>                         NA
##  4 Afghan~ Asia       1967    34.0 1.15e7      836. <NA>                         NA
##  5 Afghan~ Asia       1972    36.1 1.31e7      740. AFG                         291
##  6 Afghan~ Asia       1977    38.4 1.49e7      786. AFG                         262.
##  7 Afghan~ Asia       1982    39.9 1.29e7      978. AFG                         231.
##  8 Afghan~ Asia       1987    40.8 1.39e7      852. AFG                         198.
##  9 Afghan~ Asia       1992    41.7 1.63e7      649. AFG                         166.
## 10 Afghan~ Asia       1997    41.8 2.22e7      635. AFG                         142.
## # i 1,694 more rows
```

an alternative (that does something different) is inner_join()

```
gapminder |>
  inner_join(mortality_long)
```

```
## Joining with ‘by = join_by(country, year)‘
```

```
## # A tibble: 1,048 x 8
##    country continent  year lifeExp    pop gdpPercap country_code child_mortality
##    <chr>   <fct>     <int>   <dbl>  <int>     <dbl> <chr>                  <dbl>
##  1 Afghan~ Asia       1972    36.1 1.31e7      740. AFG                    291
##  2 Afghan~ Asia       1977    38.4 1.49e7      786. AFG                    262.
##  3 Afghan~ Asia       1982    39.9 1.29e7      978. AFG                    231.
##  4 Afghan~ Asia       1987    40.8 1.39e7      852. AFG                    198.
##  5 Afghan~ Asia       1992    41.7 1.63e7      649. AFG                    166.
##  6 Afghan~ Asia       1997    41.8 2.22e7      635. AFG                    142.
##  7 Afghan~ Asia       2002    42.1 2.53e7      727. AFG                    121.
##  8 Afghan~ Asia       2007    43.8 3.19e7      975. AFG                     99.9
##  9 Albania Europe     1972    67.7 2.26e6     3313. ALB                     NA
## 10 Albania Europe     1977    68.9 2.51e6     3533. ALB                     NA
## # i 1,038 more rows
```