HW_6_R

Group_GA

2022-10-28

The goal of tidyr is to help you create tidy data. Tidy data is data where: Every column is variable. Every row is an observation. Every cell is a single value.

Exercise 1

Tidy the data frame ex0724 from the Sleuth3 package. You can read about this data frame by typing help(ex0724) after loading Sleuth3.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(tidyr)
```

```
Sleuth3::ex0724
```

```
##    Year Denmark Netherlands Canada    USA
## 1  1950  0.5120      0.5160     NA     NA
## 2  1951  0.5174      0.5158     NA     NA
## 3  1952  0.5151      0.5158     NA     NA
## 4  1953  0.5175      0.5156     NA     NA
## 5  1954  0.5148      0.5157     NA     NA
## 6  1955  0.5169      0.5130     NA     NA
## 7  1956  0.5153      0.5150     NA     NA
## 8  1957  0.5161      0.5147     NA     NA
## 9  1958  0.5150      0.5139     NA     NA
## 10 1959  0.5139      0.5125     NA     NA
## 11 1960  0.5121      0.5135     NA     NA
## 12 1961  0.5125      0.5122     NA     NA
## 13 1962  0.5122      0.5121     NA     NA
## 14 1963  0.5132      0.5141     NA     NA
## 15 1964  0.5160      0.5143     NA     NA
## 16 1965  0.5148      0.5141     NA     NA
## 17 1966  0.5142      0.5129     NA     NA
## 18 1967  0.5135      0.5135     NA     NA
## 19 1968  0.5164      0.5116     NA     NA
```

```
## 20 1969  0.5171      0.5135     NA     NA
## 21 1970  0.5140      0.5120 0.5147 0.5134
## 22 1971  0.5170      0.5134 0.5153 0.5126
## 23 1972  0.5126      0.5112 0.5148 0.5125
## 24 1973  0.5133      0.5115 0.5149 0.5128
## 25 1974  0.5127      0.5132 0.5141 0.5133
## 26 1975  0.5108      0.5122 0.5136 0.5132
## 27 1976  0.5169      0.5148 0.5135 0.5128
## 28 1977  0.5144      0.5135 0.5145 0.5128
## 29 1978  0.5140      0.5126 0.5124 0.5129
## 30 1979  0.5141      0.5123 0.5146 0.5127
## 31 1980  0.5125      0.5128 0.5136 0.5129
## 32 1981  0.5108      0.5107 0.5133 0.5126
## 33 1982  0.5141      0.5128 0.5128 0.5123
## 34 1983  0.5117      0.5113 0.5145 0.5127
## 35 1984  0.5132      0.5132 0.5137 0.5122
## 36 1985  0.5111      0.5111 0.5144 0.5126
## 37 1986  0.5142      0.5087 0.5123 0.5122
## 38 1987  0.5173      0.5136 0.5120 0.5120
## 39 1988  0.5155      0.5117 0.5122 0.5121
## 40 1989  0.5132      0.5096 0.5123 0.5120
## 41 1990  0.5145      0.5132 0.5136 0.5120
## 42 1991  0.5131      0.5114     NA     NA
## 43 1992  0.5143      0.5129     NA     NA
## 44 1993  0.5140      0.5116     NA     NA
## 45 1994  0.5116      0.5128     NA     NA
```

```
library(Sleuth3)
data("ex0724")
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(readr)
```

```
ex0724 %>%
  tidyr::gather(Denmark:USA, key = "country", value = "proportion", na.rm = TRUE) -> clean_ex0724
```

```
clean_ex0724
```

```
##     Year     country proportion
## 1   1950     Denmark     0.5120
## 2   1951     Denmark     0.5174
## 3   1952     Denmark     0.5151
## 4   1953     Denmark     0.5175
## 5   1954     Denmark     0.5148
## 6   1955     Denmark     0.5169
## 7   1956     Denmark     0.5153
```

```
## 8   1957     Denmark     0.5161
## 9   1958     Denmark     0.5150
## 10  1959     Denmark     0.5139
## 11  1960     Denmark     0.5121
## 12  1961     Denmark     0.5125
## 13  1962     Denmark     0.5122
## 14  1963     Denmark     0.5132
## 15  1964     Denmark     0.5160
## 16  1965     Denmark     0.5148
## 17  1966     Denmark     0.5142
## 18  1967     Denmark     0.5135
## 19  1968     Denmark     0.5164
## 20  1969     Denmark     0.5171
## 21  1970     Denmark     0.5140
## 22  1971     Denmark     0.5170
## 23  1972     Denmark     0.5126
## 24  1973     Denmark     0.5133
## 25  1974     Denmark     0.5127
## 26  1975     Denmark     0.5108
## 27  1976     Denmark     0.5169
## 28  1977     Denmark     0.5144
## 29  1978     Denmark     0.5140
## 30  1979     Denmark     0.5141
## 31  1980     Denmark     0.5125
## 32  1981     Denmark     0.5108
## 33  1982     Denmark     0.5141
## 34  1983     Denmark     0.5117
## 35  1984     Denmark     0.5132
## 36  1985     Denmark     0.5111
## 37  1986     Denmark     0.5142
## 38  1987     Denmark     0.5173
## 39  1988     Denmark     0.5155
## 40  1989     Denmark     0.5132
## 41  1990     Denmark     0.5145
## 42  1991     Denmark     0.5131
## 43  1992     Denmark     0.5143
## 44  1993     Denmark     0.5140
## 45  1994     Denmark     0.5116
## 46  1950 Netherlands     0.5160
## 47  1951 Netherlands     0.5158
## 48  1952 Netherlands     0.5158
## 49  1953 Netherlands     0.5156
## 50  1954 Netherlands     0.5157
## 51  1955 Netherlands     0.5130
## 52  1956 Netherlands     0.5150
## 53  1957 Netherlands     0.5147
## 54  1958 Netherlands     0.5139
## 55  1959 Netherlands     0.5125
## 56  1960 Netherlands     0.5135
## 57  1961 Netherlands     0.5122
## 58  1962 Netherlands     0.5121
## 59  1963 Netherlands     0.5141
## 60  1964 Netherlands     0.5143
## 61  1965 Netherlands     0.5141
```

```
## 62  1966 Netherlands     0.5129
## 63  1967 Netherlands     0.5135
## 64  1968 Netherlands     0.5116
## 65  1969 Netherlands     0.5135
## 66  1970 Netherlands     0.5120
## 67  1971 Netherlands     0.5134
## 68  1972 Netherlands     0.5112
## 69  1973 Netherlands     0.5115
## 70  1974 Netherlands     0.5132
## 71  1975 Netherlands     0.5122
## 72  1976 Netherlands     0.5148
## 73  1977 Netherlands     0.5135
## 74  1978 Netherlands     0.5126
## 75  1979 Netherlands     0.5123
## 76  1980 Netherlands     0.5128
## 77  1981 Netherlands     0.5107
## 78  1982 Netherlands     0.5128
## 79  1983 Netherlands     0.5113
## 80  1984 Netherlands     0.5132
## 81  1985 Netherlands     0.5111
## 82  1986 Netherlands     0.5087
## 83  1987 Netherlands     0.5136
## 84  1988 Netherlands     0.5117
## 85  1989 Netherlands     0.5096
## 86  1990 Netherlands     0.5132
## 87  1991 Netherlands     0.5114
## 88  1992 Netherlands     0.5129
## 89  1993 Netherlands     0.5116
## 90  1994 Netherlands     0.5128
## 111 1970      Canada     0.5147
## 112 1971      Canada     0.5153
## 113 1972      Canada     0.5148
## 114 1973      Canada     0.5149
## 115 1974      Canada     0.5141
## 116 1975      Canada     0.5136
## 117 1976      Canada     0.5135
## 118 1977      Canada     0.5145
## 119 1978      Canada     0.5124
## 120 1979      Canada     0.5146
## 121 1980      Canada     0.5136
## 122 1981      Canada     0.5133
## 123 1982      Canada     0.5128
## 124 1983      Canada     0.5145
## 125 1984      Canada     0.5137
## 126 1985      Canada     0.5144
## 127 1986      Canada     0.5123
## 128 1987      Canada     0.5120
## 129 1988      Canada     0.5122
## 130 1989      Canada     0.5123
## 131 1990      Canada     0.5136
## 156 1970         USA     0.5134
## 157 1971         USA     0.5126
## 158 1972         USA     0.5125
## 159 1973         USA     0.5128
```

```
## 160 1974         USA     0.5133
## 161 1975         USA     0.5132
## 162 1976         USA     0.5128
## 163 1977         USA     0.5128
## 164 1978         USA     0.5129
## 165 1979         USA     0.5127
## 166 1980         USA     0.5129
## 167 1981         USA     0.5126
## 168 1982         USA     0.5123
## 169 1983         USA     0.5127
## 170 1984         USA     0.5122
## 171 1985         USA     0.5126
## 172 1986         USA     0.5122
## 173 1987         USA     0.5120
## 174 1988         USA     0.5121
## 175 1989         USA     0.5120
## 176 1990         USA     0.5120
```

Exercise 2

Load in and tidy the tb data frame: https://dcgerard.github.io/stat_412_612/data/tb.csv The column names specify both the sex (m = male, f = female) and age range (04 = 0 to 4, 514 = 5 to 14, 014 = 0 to 14, 1524 = 15 to 24, 2534 = 25 to 34, 3544 = 35 to 44, 4554 = 45 to 54, 4464 = 55 to 64, 65 =>=65, u = unknown).

The values in the cells are counts.Save the tidied data in the output folder.

```
tb<-read_csv("https://dcgerard.github.io/stat_412_612/data/tb.csv")
```

```
## Rows: 5769 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): iso2
## dbl (21): year, m04, m514, m014, m1524, m2534, m3544, m4554, m5564, m65, mu,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tb
```

```
## # A tibble: 5,769 x 22
##    iso2   year   m04  m514  m014 m1524 m2534 m3544 m4554 m5564   m65    mu   f04
##    <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 AD     1989    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##  2 AD     1990    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##  3 AD     1991    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##  4 AD     1992    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##  5 AD     1993    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##  6 AD     1994    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
##  7 AD     1996    NA    NA     0     0     0     4     1     0     0    NA    NA
##  8 AD     1997    NA    NA     0     0     1     2     2     1     6    NA    NA
##  9 AD     1998    NA    NA     0     0     0     1     0     0     0    NA    NA
## 10 AD     1999    NA    NA     0     0     0     1     1     0     0    NA    NA
## # ... with 5,759 more rows, and 9 more variables: f514 <dbl>, f014 <dbl>,
## #   f1524 <dbl>, f2534 <dbl>, f3544 <dbl>, f4554 <dbl>, f5564 <dbl>, f65 <dbl>,
## #   fu <dbl>
```

```
tb %>%
  tidyr::gather(-iso2, -year, key = "sex_age", value = "counts", na.rm = TRUE) %>%
  tidyr::separate(col = sex_age, into = c("sex", "age"), sep = 1) ->
  tb2
tb2
```

```
## # A tibble: 35,750 x 5
##    iso2   year sex   age   counts
##    <chr> <dbl> <chr> <chr>  <dbl>
##  1 AD     2005 m     04         0
##  2 AD     2006 m     04         0
##  3 AD     2008 m     04         0
##  4 AE     2006 m     04         0
##  5 AE     2007 m     04         0
##  6 AE     2008 m     04         0
##  7 AG     2007 m     04         0
##  8 AL     2005 m     04         0
##  9 AL     2006 m     04         1
## 10 AL     2007 m     04         0
## # ... with 35,740 more rows
```

Exercise 3

Load in and tidy the wine data frame: https://dcgerard.github.io/stat_412_612/data/wine.csv Save the tidied data in the output folder.

```
wine <- read_csv2("https://dcgerard.github.io/stat_412_612/data/wine.csv")
```

```
## i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
## Rows: 2 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr  (1): measure
## dbl (18): Norway, Scotland, England, Ireland, Finland, Canada, UnitedStates,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
wine
```

```
## # A tibble: 2 x 19
##   measure  Norway Scotl~1 England Ireland Finland Canada Unite~2 Nethe~3 NewZe~4
##   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 wine        2.8     3.2     3.2     3.4     4.3    4.9     5.1     5.2     5.9
## 2 mortali~    6.2     9       7.1     6.8    10.2    7.8     9.3     5.9     8.9
## # ... with 9 more variables: Denmark <dbl>, Sweden <dbl>, Australia <dbl>,
## #   Belgium <dbl>, Germany <dbl>, Austria <dbl>, Switzerland <dbl>,
## #   Italy <dbl>, France <dbl>, and abbreviated variable names 1: Scotland,
## #   2: UnitedStates, 3: Netherlands, 4: NewZealand
```

```
wine %>%
  tidyr::gather(-measure, key = "country", value = "value") %>%
  tidyr::spread(key = measure, value = value) ->
  wine_clean
wine_clean
```

```
## # A tibble: 18 x 3
##    country      mortality  wine
##    <chr>            <dbl> <dbl>
##  1 Australia          9.1   8.3
##  2 Austria            4.7  25.1
##  3 Belgium            5.1  12.6
##  4 Canada             7.8   4.9
##  5 Denmark            5.5   5.9
##  6 England            7.1   3.2
##  7 Finland           10.2   4.3
##  8 France             2.1  75.9
##  9 Germany            4.7  15.1
## 10 Ireland            6.8   3.4
## 11 Italy              3.2  75.9
## 12 Netherlands        5.9   5.2
## 13 NewZealand         8.9   5.9
## 14 Norway             6.2   2.8
## 15 Scotland           9     3.2
## 16 Sweden             7.1   6.6
## 17 Switzerland        3.1  33.1
## 18 UnitedStates       9.3   5.1
```