Data Tidying Using tidyr

Yogindra Raghav

October 26, 2018

```
library(dplyr)

library(tidyr)

pew <- tbl_df(read.csv("C:/Users/Yogindra Raghav/Downloads/pew.csv",
stringsAsFactors = FALSE, check.names = FALSE))



pew %>%
  gather(key = income, value = frequency, '<$10k':'$10-20k':'$20-30k':'$30-
40k':'$40-50k':'$50-75k':'$75-100k':'$100-150k':'>150k':"Don't know/refused")

## # A tibble: 180 x 3
##    religion                income frequency
##    <chr>                   <chr>      <int>
##  1 Agnostic                <$10k         27
##  2 Atheist                 <$10k         12
##  3 Buddhist                <$10k         27
##  4 Catholic                <$10k        418
##  5 Donâ□□t know/refused    <$10k         15
##  6 Evangelical Prot        <$10k        575
##  7 Hindu                   <$10k          1
##  8 Historically Black Prot <$10k        228
##  9 Jehovah's Witness       <$10k         20
## 10 Jewish                  <$10k         19
## # ... with 170 more rows

tidy4b <- table4b %>%
  gather(key = year, value = population, '1999':'2000')
tidy4b

## # A tibble: 6 x 3
##   country     year  population
##   <chr>       <chr>      <int>
## 1 Afghanistan 1999    19987071
## 2 Brazil      1999   172006362
## 3 China       1999  1272915272
## 4 Afghanistan 2000    20595360
## 5 Brazil      2000   174504898
## 6 China       2000  1280428583
```

```
tidy4a <- table4a %>%
  gather(key = year, value = case, '1999':'2000')

tidy4a %>% left_join(tidy4b) %>%
arrange(country)
```

```
## Joining, by = c("country", "year")
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## # A tibble: 6 x 4
##   country     year    case population
##   <chr>       <chr>  <int>      <int>
## 1 Afghanistan 1999     745   19987071
## 2 Afghanistan 2000    2666   20595360
## 3 Brazil      1999   37737  172006362
## 4 Brazil      2000   80488  174504898
## 5 China       1999  212258 1272915272
## 6 China       2000  213766 1280428583
```

```
stocks <- tibble(
  year   = c(2015, 2015, 2016, 2016),
  half  = c(   1,    2,     1,    2),
  return = c(1.88, 0.59, 0.92, 0.17)
)

stocks %>%  spread( key = half, value = return)
```

```
## # A tibble: 2 x 3
##    year   `1`   `2`
##   <dbl> <dbl> <dbl>
## 1  2015  1.88  0.59
## 2  2016  0.92  0.17
```

We need to gather this data set based on “sex”. The variables are “pregnant” and “sex”.

```
pregnant <- tribble(
  ~pregnant, ~male, ~female,
  "yes",     NA,    10,
  "no",      20,    12
)


pregnant %>%
  gather(key = sex, value = n, 'male':'female')
```

```
## # A tibble: 4 x 3
##   pregnant sex        n
##   <chr>    <chr>  <dbl>
## 1 yes      male      NA
## 2 no       male      20
```

```
## 3 yes      female    10
## 4 no       female    12
```

```
table5 %>%
  unite(year, century, year, sep = "") %>%
  separate(rate, into = c("cases", "population"), sep = "/", convert = TRUE)
%>%
  separate(year, into = c("year"), convert=TRUE)
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```