TidyR Article Exercises

Matthew D. Ciaramitaro

February 21, 2018

##Tidy Data

Converting Table 4 to Table 6

```
library(foreign)
library(stringr)
library(plyr)
library(reshape2)
#knitting errors
#library(tidyverse)
suppressPackageStartupMessages(library(tidyverse))

source("xtable.r")
# Data from http://pewforum.org/Datasets/Dataset-Download.aspx

# Load data ---------------------------------------------------------------

pew <- read.spss("pew.sav")
```

```
## re-encoding from CP1252
```

```
## Warning in read.spss("pew.sav"): Undeclared level(s) 2, 3, 4, 9 added in
## variable: density3
```

```
## Warning in read.spss("pew.sav"): Duplicated levels in factor denom:
## Electronic ministries
```

```
## Warning in read.spss("pew.sav"): Undeclared level(s) 1, 2, 3, 4, 5, 6, 7,
## 8, 9, 10, 11, 12, 14, 16, 23, 33 added in variable: children
```

```
## Warning in read.spss("pew.sav"): Undeclared level(s) 18, 19, 20, 21, 22,
## 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41,
## 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60,
## 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79,
## 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96 added in
## variable: age
```

```
pew <- as.data.frame(pew)


religion <- pew[c("q16", "reltrad", "income")]
religion$reltrad <- as.character(religion$reltrad)
religion$reltrad <- str_replace(religion$reltrad, " Churches", "")
religion$reltrad <- str_replace(religion$reltrad, " Protestant", " Prot")
religion$reltrad[religion$q16 == " Atheist (do not believe in God) "] <- "Atheist"
religion$reltrad[religion$q16 == " Agnostic (not sure if there is a God) "] <- "Agnostic"
religion$reltrad <- str_trim(religion$reltrad)
religion$reltrad <- str_replace_all(religion$reltrad, " \\(.*?\\)", "")

religion$income <- c("Less than $10,000" = "<$10k",
  "10 to under $20,000" = "$10-20k",
```

```
  "20 to under $30,000" = "$20-30k",
  "30 to under $40,000" = "$30-40k",
  "40 to under $50,000" = "$40-50k",
  "50 to under $75,000" = "$50-75k",
  "75 to under $100,000" = "$75-100k",
  "100 to under $150,000" = "$100-150k",
  "$150,000 or more" = ">150k",
  "Don't know/Refused (VOL)" = "Don't know/refused")[religion$income]

religion$income <- factor(religion$income, levels = c("<$10k", "$10-20k", "$20-30k", "$30-40k", "$40-50
  "$75-100k", "$100-150k", ">150k", "Don't know/refused"))
raw <- religion %>%  select(religion = q16, income)
head(raw)
```

```
##                  religion   income
## 1             Protestant  $75-100k
## 2             Protestant   $20-30k
## 3             Protestant   $30-40k
## 4  Nothing in particular     <$10k
## 5       Jewish (Judaism)   $50-75k
## 6       Jewish (Judaism)   $20-30k
```

```
raw <- raw %>%
  group_by(.dots=c("religion", "income")) %>%
  summarize(Number = n()) %>%
  spread(key = income, value = Number) %>%
  arrange(religion)
# Convert into the form in which I originally saw it ------------------------

#BEGIN ASSIGNMENT
table4 <- raw[1:10, 1:7]
head(table4,10)
```

```
## # A tibble: 10 x 7
## # Groups:   religion [10]
##    religion      `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k` `$50-75k`
##    <fct>           <int>     <int>     <int>     <int>     <int>     <int>
##  1 " Protestant~    1068      1563      1869      1831      1682      2741
##  2 " Roman Cath~     418       617       732       670       638      1115
##  3 " Mormon (Ch~      29        40        48        51        56       112
##  4 " Orthodox (~      13        17        23        32        32        47
##  5 " Jewish (Ju~      19        19        25        25        30        95
##  6 " Muslim (Is~       6         7         9        10         9        23
##  7 " Buddhist "       27        21        30        34        33        58
##  8 " Hindu "           1         9         7         9        11        34
##  9 " Atheist (d~      12        27        37        52        35        70
## 10 " Agnostic (~      27        34        60        81        76       137
```

At this point we have the correct starting table. Now we will use diplyr and tidyr to tidy the table to be the same as table 6.

```
table6 <- table4 %>%   gather(key = "income",value = "freq", 2:7)
table6 <- table6 %>% arrange(religion)
head(table6,10)
```

```
## # A tibble: 10 x 3
```

```
## # Groups:   religion [2]
##    religion           income   freq
##    <fct>              <chr>   <int>
##  1 " Protestant "     <$10k    1068
##  2 " Protestant "     $10-20k  1563
##  3 " Protestant "     $20-30k  1869
##  4 " Protestant "     $30-40k  1831
##  5 " Protestant "     $40-50k  1682
##  6 " Protestant "     $50-75k  2741
##  7 " Roman Catholic " <$10k     418
##  8 " Roman Catholic " $10-20k   617
##  9 " Roman Catholic " $20-30k   732
## 10 " Roman Catholic " $30-40k   670
```

So now we have completed the table as table 6.

##Converting Table 7 to table 8

```
raw <- read_csv("billboard.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   artist.inverted = col_character(),
##   track = col_character(),
##   time = col_time(format = ""),
##   genre = col_character(),
##   date.entered = col_date(format = ""),
##   date.peaked = col_date(format = ""),
##   x66th.week = col_character(),
##   x67th.week = col_character(),
##   x68th.week = col_character(),
##   x69th.week = col_character(),
##   x70th.week = col_character(),
##   x71st.week = col_character(),
##   x72nd.week = col_character(),
##   x73rd.week = col_character(),
##   x74th.week = col_character(),
##   x75th.week = col_character(),
##   x76th.week = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
head(raw)
```

```
## # A tibble: 6 x 83
##    year artist.inverted  track        time  genre date.entered date.peaked
##   <int> <chr>            <chr>        <tim> <chr> <date>       <date>
## 1  2000 Destiny's Child  Independent~ 03:38 Rock  2000-09-23   2000-11-18
## 2  2000 Santana          Maria, Maria 04:18 Rock  2000-02-12   2000-04-08
## 3  2000 Savage Garden    I Knew I Lo~ 04:07 Rock  1999-10-23   2000-01-29
## 4  2000 Madonna          Music        03:45 Rock  2000-08-12   2000-09-16
## 5  2000 Aguilera, Chris~ Come On Ove~ 03:38 Rock  2000-08-05   2000-10-14
## 6  2000 Janet            Doesn't Rea~ 04:17 Rock  2000-06-17   2000-08-26
## # ... with 76 more variables: x1st.week <int>, x2nd.week <int>,
## #   x3rd.week <int>, x4th.week <int>, x5th.week <int>, x6th.week <int>,
```

```
## #   x7th.week <int>, x8th.week <int>, x9th.week <int>, x10th.week <int>,
## #   x11th.week <int>, x12th.week <int>, x13th.week <int>,
## #   x14th.week <int>, x15th.week <int>, x16th.week <int>,
## #   x17th.week <int>, x18th.week <int>, x19th.week <int>,
## #   x20th.week <int>, x21st.week <int>, x22nd.week <int>,
## #   x23rd.week <int>, x24th.week <int>, x25th.week <int>,
## #   x26th.week <int>, x27th.week <int>, x28th.week <int>,
## #   x29th.week <int>, x30th.week <int>, x31st.week <int>,
## #   x32nd.week <int>, x33rd.week <int>, x34th.week <int>,
## #   x35th.week <int>, x36th.week <int>, x37th.week <int>,
## #   x38th.week <int>, x39th.week <int>, x40th.week <int>,
## #   x41st.week <int>, x42nd.week <int>, x43rd.week <int>,
## #   x44th.week <int>, x45th.week <int>, x46th.week <int>,
## #   x47th.week <int>, x48th.week <int>, x49th.week <int>,
## #   x50th.week <int>, x51st.week <int>, x52nd.week <int>,
## #   x53rd.week <int>, x54th.week <int>, x55th.week <int>,
## #   x56th.week <int>, x57th.week <int>, x58th.week <int>,
## #   x59th.week <int>, x60th.week <int>, x61st.week <int>,
## #   x62nd.week <int>, x63rd.week <int>, x64th.week <int>,
## #   x65th.week <int>, x66th.week <chr>, x67th.week <chr>,
## #   x68th.week <chr>, x69th.week <chr>, x70th.week <chr>,
## #   x71st.week <chr>, x72nd.week <chr>, x73rd.week <chr>,
## #   x74th.week <chr>, x75th.week <chr>, x76th.week <chr>
```

```
table7 <- raw %>% #getraw data
          select(year, artist=artist.inverted,track, time, date.entered, starts_with("x")) %>% #grab ne
          arrange(artist) #sort table
head(table7,10)
```

```
## # A tibble: 10 x 81
##     year artist  track    time  date.entered x1st.week x2nd.week x3rd.week
##    <int> <chr>   <chr>    <tim> <date>           <int>     <int>     <int>
##  1  2000 2 Pac   Baby Do~ 04:22 2000-02-26          87        82        72
##  2  2000 2Ge+her The Har~ 03:15 2000-09-02          91        87        92
##  3  2000 3 Door~ Krypton~ 03:53 2000-04-08          81        70        68
##  4  2000 3 Door~ Loser    04:24 2000-10-21          76        76        72
##  5  2000 504 Bo~ Wobble ~ 03:35 2000-04-15          57        34        25
##  6  2000 "98\xa~ Give Me~ 03:24 2000-08-19          51        39        34
##  7  2000 A*Teens Dancing~ 03:44 2000-07-08          97        97        96
##  8  2000 Aaliyah Try Aga~ 04:03 2000-03-18          59        53        38
##  9  2000 Aaliyah I Don't~ 04:15 2000-01-29          84        62        51
## 10  2000 Adams,~ Open My~ 05:30 2000-08-26          76        76        74
## # ... with 73 more variables: x4th.week <int>, x5th.week <int>,
## #   x6th.week <int>, x7th.week <int>, x8th.week <int>, x9th.week <int>,
## #   x10th.week <int>, x11th.week <int>, x12th.week <int>,
## #   x13th.week <int>, x14th.week <int>, x15th.week <int>,
## #   x16th.week <int>, x17th.week <int>, x18th.week <int>,
## #   x19th.week <int>, x20th.week <int>, x21st.week <int>,
## #   x22nd.week <int>, x23rd.week <int>, x24th.week <int>,
## #   x25th.week <int>, x26th.week <int>, x27th.week <int>,
## #   x28th.week <int>, x29th.week <int>, x30th.week <int>,
## #   x31st.week <int>, x32nd.week <int>, x33rd.week <int>,
## #   x34th.week <int>, x35th.week <int>, x36th.week <int>,
## #   x37th.week <int>, x38th.week <int>, x39th.week <int>,
## #   x40th.week <int>, x41st.week <int>, x42nd.week <int>,
```

```
## #   x43rd.week <int>, x44th.week <int>, x45th.week <int>,
## #   x46th.week <int>, x47th.week <int>, x48th.week <int>,
## #   x49th.week <int>, x50th.week <int>, x51st.week <int>,
## #   x52nd.week <int>, x53rd.week <int>, x54th.week <int>,
## #   x55th.week <int>, x56th.week <int>, x57th.week <int>,
## #   x58th.week <int>, x59th.week <int>, x60th.week <int>,
## #   x61st.week <int>, x62nd.week <int>, x63rd.week <int>,
## #   x64th.week <int>, x65th.week <int>, x66th.week <chr>,
## #   x67th.week <chr>, x68th.week <chr>, x69th.week <chr>,
## #   x70th.week <chr>, x71st.week <chr>, x72nd.week <chr>,
## #   x73rd.week <chr>, x74th.week <chr>, x75th.week <chr>, x76th.week <chr>
```

Now table 7 is formatted correctly

```
table8 <- table7  %>% gather(key="week", value = "rank", -year, -artist, -track, -time, -date.entered) %
        select(year, artist, time, track, date = date.entered, week, rank )
head(table8)
```

```
## # A tibble: 6 x 7
##    year artist       time   track                  date       week   rank
##   <int> <chr>        <time> <chr>                  <date>     <chr>  <chr>
## 1  2000 2 Pac        04:22  Baby Don't Cry (Keep ~ 2000-02-26 x1st.~ 87
## 2  2000 2Ge+her      03:15  The Hardest Part Of B~ 2000-09-02 x1st.~ 91
## 3  2000 3 Doors Down 03:53  Kryptonite             2000-04-08 x1st.~ 81
## 4  2000 3 Doors Down 04:24  Loser                  2000-10-21 x1st.~ 76
## 5  2000 504 Boyz     03:35  Wobble Wobble          2000-04-15 x1st.~ 57
## 6  2000 "98\xa1"     03:24  Give Me Just One Nigh~ 2000-08-19 x1st.~ 51
```

```
table8 <- table8 %>% arrange(track) %>%
        filter(!is.na(rank)) %>%
        mutate(week = as.integer(str_extract(week, "[0-9]+"))) %>%
        arrange(artist, track) %>%
        mutate(date = date + (week-1)*7 ) %>%
        mutate(rank = as.integer(rank))
options(tibble.print_max = 15)
head(table8, 15)
```

```
## # A tibble: 15 x 7
##     year artist       time   track                  date        week  rank
##    <int> <chr>        <time> <chr>                  <date>     <int> <int>
##  1  2000 2 Pac        04:22  Baby Don't Cry (Keep ~ 2000-02-26     1    87
##  2  2000 2 Pac        04:22  Baby Don't Cry (Keep ~ 2000-03-04     2    82
##  3  2000 2 Pac        04:22  Baby Don't Cry (Keep ~ 2000-03-11     3    72
##  4  2000 2 Pac        04:22  Baby Don't Cry (Keep ~ 2000-03-18     4    77
##  5  2000 2 Pac        04:22  Baby Don't Cry (Keep ~ 2000-03-25     5    87
##  6  2000 2 Pac        04:22  Baby Don't Cry (Keep ~ 2000-04-01     6    94
##  7  2000 2 Pac        04:22  Baby Don't Cry (Keep ~ 2000-04-08     7    99
##  8  2000 2Ge+her      03:15  The Hardest Part Of B~ 2000-09-02     1    91
##  9  2000 2Ge+her      03:15  The Hardest Part Of B~ 2000-09-09     2    87
## 10  2000 2Ge+her      03:15  The Hardest Part Of B~ 2000-09-16     3    92
## 11  2000 3 Doors Down 03:53  Kryptonite             2000-04-08     1    81
## 12  2000 3 Doors Down 03:53  Kryptonite             2000-04-15     2    70
## 13  2000 3 Doors Down 03:53  Kryptonite             2000-04-22     3    68
## 14  2000 3 Doors Down 03:53  Kryptonite             2000-04-29     4    67
## 15  2000 3 Doors Down 03:53  Kryptonite             2000-05-06     5    66
```