R for Data Science: TB Case Study

Monica Buczynski

8/18/2020

Note: The purpose of this document is to showcase a sample of skills that I learned in *R for Data Science* (chapter: Tidy Data) by Garrett Grolemund and Hadley Wickham. Particularly, I conduct a case study concerning tuberculosis. All scripts were taken from https://r4ds.had.co.nz/tidy-data.html and https://jrnold.github.io/r4ds-exercise-solutions/index.html. The code for each exercise was studied carefully for understanding and then was retyped manually into R to maximize the learning experience; however, many of the original scripts were altered for further experimentation and presentation aesthetics.

```
# View first 10 rows of data set and all column names.

who
```

```
## # A tibble: 7,240 x 60
##    country iso2  iso3   year new_sp_m014 new_sp_m1524 new_sp_m2534 new_sp_m3544
##    <chr>   <chr> <chr> <int>       <int>        <int>        <int>        <int>
##  1 Afghan~ AF    AFG    1980          NA           NA           NA           NA
##  2 Afghan~ AF    AFG    1981          NA           NA           NA           NA
##  3 Afghan~ AF    AFG    1982          NA           NA           NA           NA
##  4 Afghan~ AF    AFG    1983          NA           NA           NA           NA
##  5 Afghan~ AF    AFG    1984          NA           NA           NA           NA
##  6 Afghan~ AF    AFG    1985          NA           NA           NA           NA
##  7 Afghan~ AF    AFG    1986          NA           NA           NA           NA
##  8 Afghan~ AF    AFG    1987          NA           NA           NA           NA
##  9 Afghan~ AF    AFG    1988          NA           NA           NA           NA
## 10 Afghan~ AF    AFG    1989          NA           NA           NA           NA
## # ... with 7,230 more rows, and 52 more variables: new_sp_m4554 <int>,
## #   new_sp_m5564 <int>, new_sp_m65 <int>, new_sp_f014 <int>,
## #   new_sp_f1524 <int>, new_sp_f2534 <int>, new_sp_f3544 <int>,
## #   new_sp_f4554 <int>, new_sp_f5564 <int>, new_sp_f65 <int>,
## #   new_sn_m014 <int>, new_sn_m1524 <int>, new_sn_m2534 <int>,
## #   new_sn_m3544 <int>, new_sn_m4554 <int>, new_sn_m5564 <int>,
## #   new_sn_m65 <int>, new_sn_f014 <int>, new_sn_f1524 <int>,
## #   new_sn_f2534 <int>, new_sn_f3544 <int>, new_sn_f4554 <int>,
## #   new_sn_f5564 <int>, new_sn_f65 <int>, new_ep_m014 <int>,
## #   new_ep_m1524 <int>, new_ep_m2534 <int>, new_ep_m3544 <int>,
## #   new_ep_m4554 <int>, new_ep_m5564 <int>, new_ep_m65 <int>,
## #   new_ep_f014 <int>, new_ep_f1524 <int>, new_ep_f2534 <int>,
## #   new_ep_f3544 <int>, new_ep_f4554 <int>, new_ep_f5564 <int>,
## #   new_ep_f65 <int>, newrel_m014 <int>, newrel_m1524 <int>,
## #   newrel_m2534 <int>, newrel_m3544 <int>, newrel_m4554 <int>,
## #   newrel_m5564 <int>, newrel_m65 <int>, newrel_f014 <int>,
## #   newrel_f1524 <int>, newrel_f2534 <int>, newrel_f3544 <int>,
## #   newrel_f4554 <int>, newrel_f5564 <int>, newrel_f65 <int>
```

```
(who1 <- who %>%
  pivot_longer(
    cols = new_sp_m014:newrel_f65, # aggregating all of the columns into the generic grouping "key" ->
    names_to = "key",
    values_to = "cases",
    values_drop_na = TRUE # focus on the values that we have present
  ))
```

```
## # A tibble: 76,046 x 6
##    country     iso2  iso3   year key          cases
##    <chr>       <chr> <chr> <int> <chr>        <int>
##  1 Afghanistan AF    AFG    1997 new_sp_m014      0
##  2 Afghanistan AF    AFG    1997 new_sp_m1524    10
##  3 Afghanistan AF    AFG    1997 new_sp_m2534     6
##  4 Afghanistan AF    AFG    1997 new_sp_m3544     3
##  5 Afghanistan AF    AFG    1997 new_sp_m4554     5
##  6 Afghanistan AF    AFG    1997 new_sp_m5564     2
##  7 Afghanistan AF    AFG    1997 new_sp_m65       0
##  8 Afghanistan AF    AFG    1997 new_sp_f014      5
```

```
##  9 Afghanistan AF    AFG    1997 new_sp_f1524    38
## 10 Afghanistan AF    AFG    1997 new_sp_f2534    36
## # ... with 76,036 more rows
```

```
# Count all of the values in the new "key" column

who1 %>%
  count(key)
```

```
## # A tibble: 56 x 2
##    key              n
##    <chr>        <int>
##  1 new_ep_f014   1032
##  2 new_ep_f1524  1021
##  3 new_ep_f2534  1021
##  4 new_ep_f3544  1021
##  5 new_ep_f4554  1017
##  6 new_ep_f5564  1017
##  7 new_ep_f65    1014
##  8 new_ep_m014   1038
##  9 new_ep_m1524  1026
## 10 new_ep_m2534  1020
## # ... with 46 more rows
```

```
# Make all variable names consistent

(who2 <- who1 %>%
  mutate(names_from = stringr::str_replace(key, "newrel", "new_rel")))
```

```
## # A tibble: 76,046 x 7
##    country     iso2  iso3   year key          cases names_from
##    <chr>       <chr> <chr> <int> <chr>        <int> <chr>
##  1 Afghanistan AF    AFG    1997 new_sp_m014      0 new_sp_m014
##  2 Afghanistan AF    AFG    1997 new_sp_m1524    10 new_sp_m1524
##  3 Afghanistan AF    AFG    1997 new_sp_m2534     6 new_sp_m2534
##  4 Afghanistan AF    AFG    1997 new_sp_m3544     3 new_sp_m3544
##  5 Afghanistan AF    AFG    1997 new_sp_m4554     5 new_sp_m4554
##  6 Afghanistan AF    AFG    1997 new_sp_m5564     2 new_sp_m5564
##  7 Afghanistan AF    AFG    1997 new_sp_m65       0 new_sp_m65
##  8 Afghanistan AF    AFG    1997 new_sp_f014      5 new_sp_f014
##  9 Afghanistan AF    AFG    1997 new_sp_f1524    38 new_sp_f1524
## 10 Afghanistan AF    AFG    1997 new_sp_f2534    36 new_sp_f2534
## # ... with 76,036 more rows
```

```
# Separate the values in each code with two passes of separate().

(who3 <- who2 %>%
  separate(key, c("new", "type", "sexage"), sep = "_"))
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 2580 rows [243,
## 244, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 903,
## 904, 905, 906, ...].
```

```
## # A tibble: 76,046 x 9
##    country     iso2  iso3   year new   type  sexage cases names_from
##    <chr>       <chr> <chr> <int> <chr> <chr> <chr>  <int> <chr>
##  1 Afghanistan AF    AFG    1997 new   sp    m014       0 new_sp_m014
```

```
##  2 Afghanistan AF    AFG    1997 new   sp    m1524     10 new_sp_m1524
##  3 Afghanistan AF    AFG    1997 new   sp    m2534      6 new_sp_m2534
##  4 Afghanistan AF    AFG    1997 new   sp    m3544      3 new_sp_m3544
##  5 Afghanistan AF    AFG    1997 new   sp    m4554      5 new_sp_m4554
##  6 Afghanistan AF    AFG    1997 new   sp    m5564      2 new_sp_m5564
##  7 Afghanistan AF    AFG    1997 new   sp    m65        0 new_sp_m65
##  8 Afghanistan AF    AFG    1997 new   sp    f014       5 new_sp_f014
##  9 Afghanistan AF    AFG    1997 new   sp    f1524     38 new_sp_f1524
## 10 Afghanistan AF    AFG    1997 new   sp    f2534     36 new_sp_f2534
## # ... with 76,036 more rows
```

```
# Drop "iso2", "iso3" because they are redundant. Drop "new" because it is constant throughout the enti

(who4 <- who3 %>%
  select(-new, -iso2, -iso3))
```

```
## # A tibble: 76,046 x 6
##    country      year type  sexage cases names_from
##    <chr>       <int> <chr> <chr>  <int> <chr>
##  1 Afghanistan  1997 sp    m014       0 new_sp_m014
##  2 Afghanistan  1997 sp    m1524     10 new_sp_m1524
##  3 Afghanistan  1997 sp    m2534      6 new_sp_m2534
##  4 Afghanistan  1997 sp    m3544      3 new_sp_m3544
##  5 Afghanistan  1997 sp    m4554      5 new_sp_m4554
##  6 Afghanistan  1997 sp    m5564      2 new_sp_m5564
##  7 Afghanistan  1997 sp    m65        0 new_sp_m65
##  8 Afghanistan  1997 sp    f014       5 new_sp_f014
##  9 Afghanistan  1997 sp    f1524     38 new_sp_f1524
## 10 Afghanistan  1997 sp    f2534     36 new_sp_f2534
## # ... with 76,036 more rows
```

```
# Separate sex and age by splitting after the first character

(who5 <- who4 %>%
    separate(sexage, c("sex", "age"), sep = 1))
```

```
## # A tibble: 76,046 x 7
##    country      year type  sex   age   cases names_from
##    <chr>       <int> <chr> <chr> <chr> <int> <chr>
##  1 Afghanistan  1997 sp    m     014       0 new_sp_m014
##  2 Afghanistan  1997 sp    m     1524     10 new_sp_m1524
##  3 Afghanistan  1997 sp    m     2534      6 new_sp_m2534
##  4 Afghanistan  1997 sp    m     3544      3 new_sp_m3544
##  5 Afghanistan  1997 sp    m     4554      5 new_sp_m4554
##  6 Afghanistan  1997 sp    m     5564      2 new_sp_m5564
##  7 Afghanistan  1997 sp    m     65        0 new_sp_m65
##  8 Afghanistan  1997 sp    f     014       5 new_sp_f014
##  9 Afghanistan  1997 sp    f     1524     38 new_sp_f1524
## 10 Afghanistan  1997 sp    f     2534     36 new_sp_f2534
## # ... with 76,036 more rows
```

```r
# Same exercise, but built in an complex pipe instead of individual pipes

who %>%
  pivot_longer(
    cols = new_sp_m014:newrel_f65,
    names_to = "key",
    values_to = "cases",
    values_drop_na = TRUE
  ) %>%
  mutate (
    key = stringr::str_replace(key, "newrel", "new_rel")
  ) %>%
  separate(key, c("new", "var", "sexage")) %>%
  select(-new, -iso2, -iso3) %>%
  separate(sexage, c("sex", "age"), sep = 1)
```

```
## # A tibble: 76,046 x 6
##    country      year var   sex   age   cases
##    <chr>       <int> <chr> <chr> <chr> <int>
##  1 Afghanistan  1997 sp    m     014       0
##  2 Afghanistan  1997 sp    m     1524     10
##  3 Afghanistan  1997 sp    m     2534      6
##  4 Afghanistan  1997 sp    m     3544      3
##  5 Afghanistan  1997 sp    m     4554      5
##  6 Afghanistan  1997 sp    m     5564      2
##  7 Afghanistan  1997 sp    m     65        0
##  8 Afghanistan  1997 sp    f     014       5
##  9 Afghanistan  1997 sp    f     1524     38
## 10 Afghanistan  1997 sp    f     2534     36
## # ... with 76,036 more rows
```