Assign4b

Re Francesco Ignazio

21 febbraio 2018

1

In this case study I set na.rm = TRUE just to make it easier to check that we had the correct values. Is this reasonable? Think about how missing values are represented in this dataset. Are there implicit missing values? What's the difference between an NA and zero?

Yes, it is reasonable as the missing values we removed were referring to the lack of data regarding tubercolosis cases in a given country on a given year. For this reason, the rows with the missing values were redundant as they were not giving any useful information. These missing values may have been inserted explicitily in order to avoid having implicit missing values (with the NA values inserted, every country presents an observation in every year of the study (from 1980 to 2013)). However,because these explicitily missing values are considered not important, we can turn them implicit.The main difference between NA and 0 is that 0 is an actual value saying that there were no cases of tubercolosis in a country in a given year, whereas a NA value tells us we have no information on the observation at all.

2

What happens if you neglect the mutate() step?

Neglecting the mutate() step, all the observations representing cases of relapsing would be still represented by the form `newrel_****`. If we kept this form and executed the following statement

```
who3a <- who2a %>%
  separate(key, c("new", "type", "sexage"), sep = "_")
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 2580 rows
## [73467, 73468, 73469, 73470, 73471, 73472, 73473, 73474, 73475, 73476,
## 73477, 73478, 73479, 73480, 73481, 73482, 73483, 73484, 73485, 73486, ...].
```

we would have missing values in the `sexage` variable for every observation representing cases of relapsing. That's because the above statement would find only one separator `_`, for which the variable `new` would have values equal to `newrel` and the variable `type` would have the values related to `sexage`, as we can see below.

```
who3n <- who2 %>%
  separate(key, c("new", "type", "sexage"), sep = "_")

filter(who3n, new == "newrel") %>% head()
```

```
## # A tibble: 0 x 8
## # ... with 8 variables: country <chr>, iso2 <chr>, iso3 <chr>, year <int>,
## #   new <chr>, type <chr>, sexage <chr>, cases <int>
```

3

I claimed that iso2 and iso3 were redundant with country. Confirm this claim.

We can see below that after selecting the columns `country`, `iso2` and `iso3` and after taking only the rows with distinct observations from one another, if we group by country, ther will be no group with more than one element. This means that every value of “iso2” and “iso3” identifies one and one country only and hence, is redundant.

```
select(who3, country, iso2, iso3) %>%
  distinct() %>%
  group_by(country) %>%
  filter(n() > 1)
```

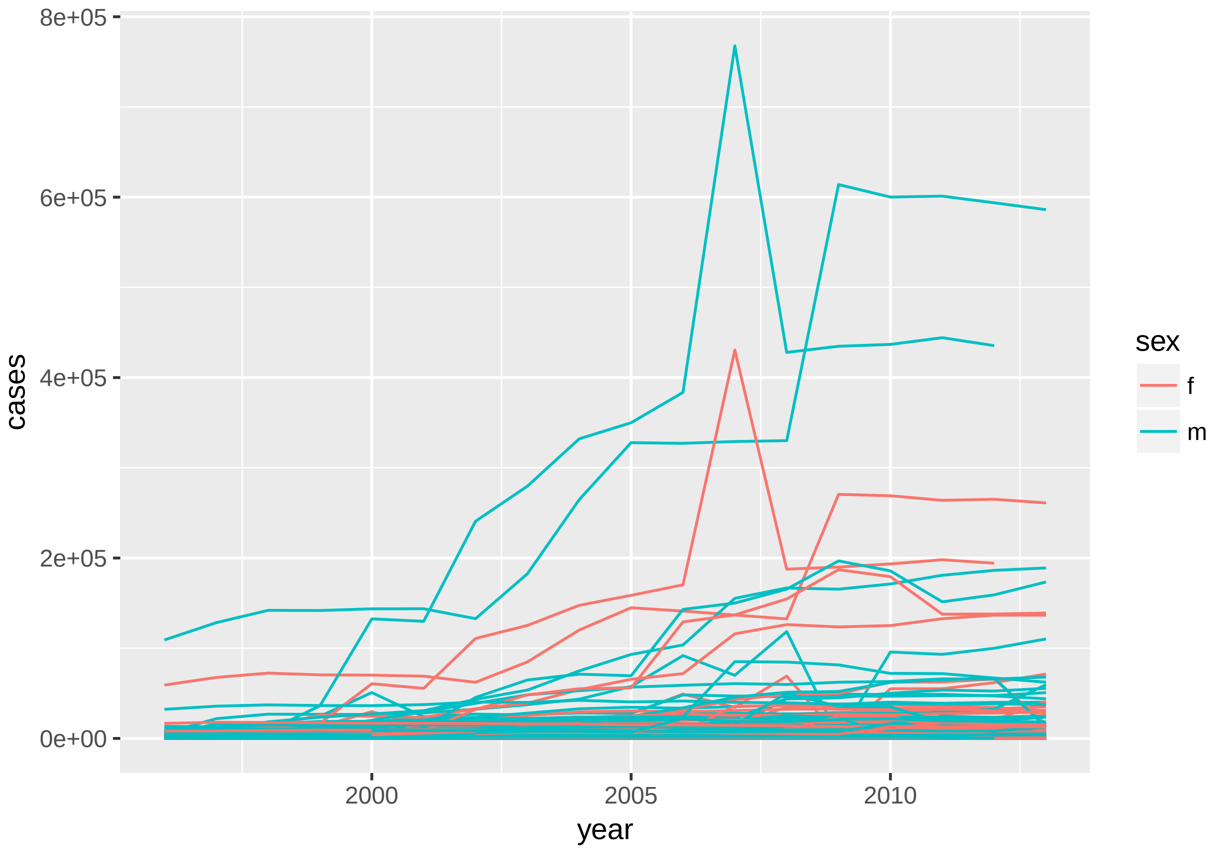
```
## # A tibble: 0 x 3
## # Groups:   country [0]
## # ... with 3 variables: country <chr>, iso2 <chr>, iso3 <chr>
```

4

For each country, year, and sex compute the total number of cases of TB. Make an informative visualization of the data.

We can see that most countries don’t ever reach an annual number of cases greater than 2e+05. We see that overall the number of cases in males is greater than the number of cases in females.

```
who5 %>% group_by(country, year, sex) %>%
   filter(year > 1995) %>%
   summarise(cases = sum(cases)) %>%
   unite(country_sex, country, sex, remove = FALSE) %>%
   ggplot(aes(x=year, y = cases)) + geom_line(aes(group = country_sex, color = sex))
```


8e+05
6e+05
4e+05
2e+05
0e+00
cases
year
sex
f
m