hw4.2

Yue Wu

2/24/2018

10.5 Exercises

5. What does tibble::enframe() do? When might you use it?

It converts vectors or lists to two-column data frames.

```
enframe(c(a = 5, b = 7))
```

```
## # A tibble: 2 x 2
##   name  value
##   <chr> <dbl>
## 1 a      5.00
## 2 b      7.00
```

12.6.1 Exercises

Repeat the case study

```
who1 <- who %>%
  gather(new_sp_m014:newrel_f65, key = "key", value = "cases", na.rm = TRUE)
who2 <- who1 %>%
  mutate(key = stringr::str_replace(key, "newrel", "new_rel"))
who3 <- who2 %>%
  separate(key, c("new", "type", "sexage"), sep = "_")
who4 <- who3 %>%
  select(-new, -iso2, -iso3)
who5 <- who4 %>%
  separate(sexage, c("sex", "age"), sep = 1)
who %>%
  gather(code, value, new_sp_m014:newrel_f65, na.rm = TRUE) %>%
  mutate(code = stringr::str_replace(code, "newrel", "new_rel")) %>%
  separate(code, 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   value
##  * <chr>       <int> <chr> <chr> <chr> <int>
##  1 Afghanistan  1997 sp    m     014       0
##  2 Afghanistan  1998 sp    m     014      30
##  3 Afghanistan  1999 sp    m     014       8
##  4 Afghanistan  2000 sp    m     014      52
##  5 Afghanistan  2001 sp    m     014     129
##  6 Afghanistan  2002 sp    m     014      90
##  7 Afghanistan  2003 sp    m     014     127
##  8 Afghanistan  2004 sp    m     014     139
##  9 Afghanistan  2005 sp    m     014     151
## 10 Afghanistan  2006 sp    m     014     193
## # ... with 76,036 more rows
```

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

None of the countries have multiple iso2 or iso3 codes.

```
whoex3 <- select(who3, country, iso2, iso3) %>%
           group_by(country)
n_groups(whoex3)
```

```
## [1] 219
```

```
whoex3_2 <-  whoex3 %>% group_by(country,iso2,iso3)
n_groups(whoex3_2)
```

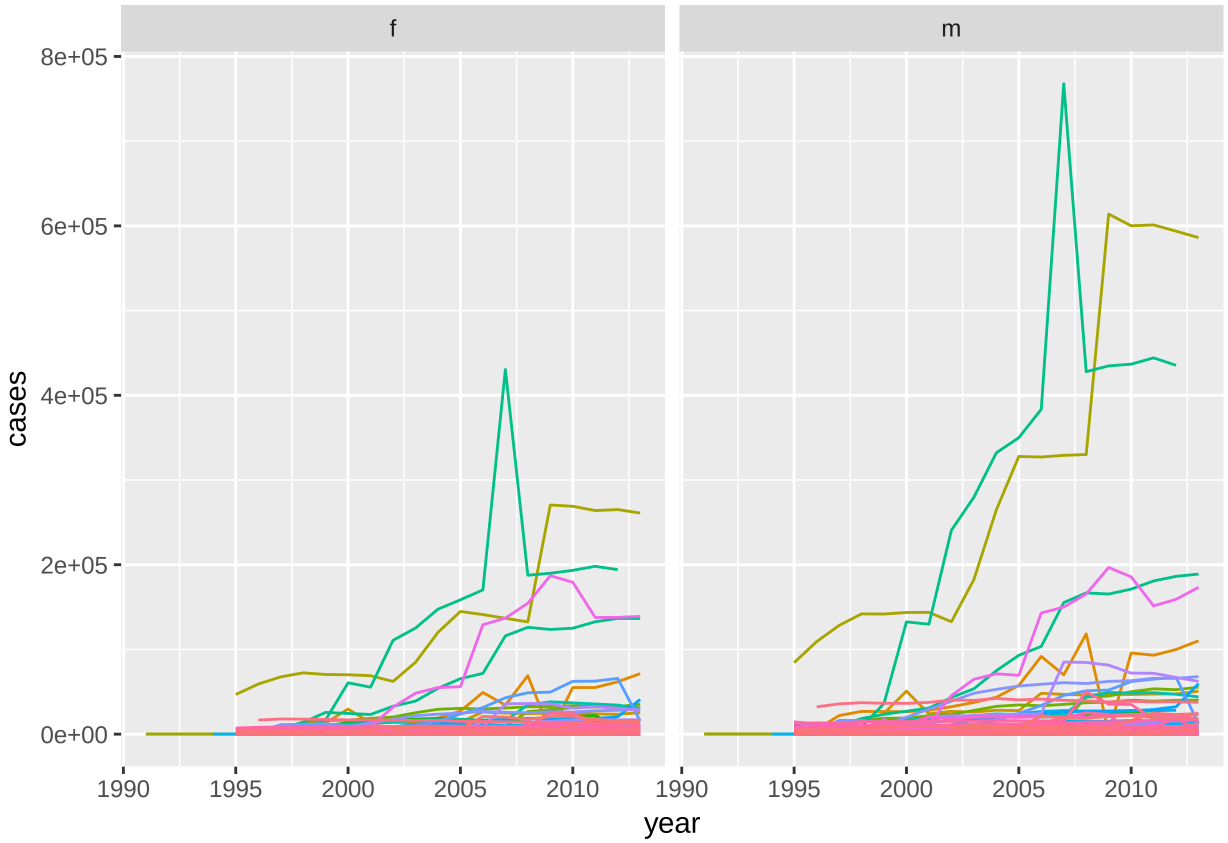
```
## [1] 219
```

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

```
whoex4 <- who5 %>%
  group_by(country, year, sex) %>%
  dplyr::summarise(cases=sum(cases))
whoex4
```

```
## # A tibble: 6,921 x 4
## # Groups:   country, year [?]
##    country      year sex   cases
##    <chr>       <int> <chr> <int>
##  1 Afghanistan  1997 f       102
##  2 Afghanistan  1997 m        26
##  3 Afghanistan  1998 f      1207
##  4 Afghanistan  1998 m       571
##  5 Afghanistan  1999 f       517
##  6 Afghanistan  1999 m       228
##  7 Afghanistan  2000 f      1751
##  8 Afghanistan  2000 m       915
##  9 Afghanistan  2001 f      3062
## 10 Afghanistan  2001 m      1577
## # ... with 6,911 more rows
```

```
whoex4  %>% filter(year>1990) %>%
  ggplot(aes(x = year, y = cases, group = country, color= country))   +geom_line()+ facet_wrap(~sex)+ t
```



Tidying: Table 4 to Table 6

Table 4

```
raw1[1:10,1:7]
```

```
##                   religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k
## 1                 Agnostic    27      34      60      81      76     137
## 2                  Atheist    12      27      37      52      35      70
## 3                 Buddhist    27      21      30      34      33      58
## 4                 Catholic   418     617     732     670     638    1116
## 5       Don’t know/refused    15      14      15      11      10      35
## 6         Evangelical Prot   575     869    1064     982     881    1486
## 7                    Hindu     1       9       7       9      11      34
## 8  Historically Black Prot   228     244     236     238     197     223
## 9           Jehovah's Witness    20      27      24      24      21      30
## 10                  Jewish    19      19      25      25      30      95
```

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.

Table 6

```
tb6 <- as.tibble(raw1)
tb6 %>% gather(-religion, key = "income", value = "freq") %>% arrange(religion) %>% head(n=10)
```

```
## # A tibble: 10 x 3
##    religion income              freq
##    <chr>    <chr>              <int>
##  1 Agnostic <$10k                 27
##  2 Agnostic $10-20k               34
##  3 Agnostic $20-30k               60
##  4 Agnostic $30-40k               81
##  5 Agnostic $40-50k               76
##  6 Agnostic $50-75k              137
##  7 Agnostic $75-100k             122
##  8 Agnostic $100-150k            109
##  9 Agnostic >150k                 84
## 10 Agnostic Don't know/refused    96
```

Table 6: The first ten rows of the tidied Pew survey dataset on income and religion. The column has been renamed to income, and value to freq.

Tidying: Table 7 to Table 8

Table 7

```r
raw[c(1:3, 6:10),1:8]
```

```
##    year         artist                   track time date.entered wk1 wk2
## 1  2000          2 Pac          Baby Don't Cry 4:22   2000-02-26  87  82
## 2  2000        2Ge+her The Hardest Part Of ... 3:15   2000-09-02  91  87
## 3  2000   3 Doors Down              Kryptonite 3:53   2000-04-08  81  70
## 6  2000           98^0 Give Me Just One Nig... 3:24   2000-08-19  51  39
## 7  2000        A*Teens           Dancing Queen 3:44   2000-07-08  97  97
## 8  2000        Aaliyah           I Don't Wanna 4:15   2000-01-29  84  62
## 9  2000        Aaliyah               Try Again 4:03   2000-03-18  59  53
## 10 2000 Adams, Yolanda           Open My Heart 5:30   2000-08-26  76  76
##    wk3
## 1   72
## 2   92
## 3   68
## 6   34
## 7   96
## 8   51
## 9   38
## 10  74
```

Table 7: The first eight Billboard top hits for 2000. Other columns not shown are $wk4, wk5, ..., wk75$.

Table 8

```r
tb7 <- as.tibble(raw)
tb8 <- tb7 %>% gather(key= "week", value= "rank", -year, -artist, -track, -time, -date.entered ) %>%
  arrange(artist) %>%
  select(year,artist, time, track, date=date.entered,week,rank) %>%
  filter(!is.na(rank))
head(tb8,n=15)
```

```
## # A tibble: 15 x 7
##     year artist       time  track                   date       week   rank
##    <int> <chr>        <chr> <chr>                   <chr>      <chr> <int>
##  1  2000 2 Pac        4:22  Baby Don't Cry          2000-02-26 wk1      87
##  2  2000 2 Pac        4:22  Baby Don't Cry          2000-02-26 wk2      82
##  3  2000 2 Pac        4:22  Baby Don't Cry          2000-02-26 wk3      72
##  4  2000 2 Pac        4:22  Baby Don't Cry          2000-02-26 wk4      77
##  5  2000 2 Pac        4:22  Baby Don't Cry          2000-02-26 wk5      87
##  6  2000 2 Pac        4:22  Baby Don't Cry          2000-02-26 wk6      94
##  7  2000 2 Pac        4:22  Baby Don't Cry          2000-02-26 wk7      99
##  8  2000 2Ge+her      3:15  The Hardest Part Of ... 2000-09-02 wk1      91
##  9  2000 2Ge+her      3:15  The Hardest Part Of ... 2000-09-02 wk2      87
## 10  2000 2Ge+her      3:15  The Hardest Part Of ... 2000-09-02 wk3      92
## 11  2000 3 Doors Down 3:53  Kryptonite              2000-04-08 wk1      81
## 12  2000 3 Doors Down 4:24  Loser                   2000-10-21 wk1      76
## 13  2000 3 Doors Down 3:53  Kryptonite              2000-04-08 wk2      70
## 14  2000 3 Doors Down 4:24  Loser                   2000-10-21 wk2      76
## 15  2000 3 Doors Down 3:53  Kryptonite              2000-04-08 wk3      68
```

Table 8: First fifteen rows of the tidied Billboard dataset. The date column does not appear in the original table, but can be computed from *date.entered* and *week*.