HW1 - Data wrangling - Προθεσμία υποβολής 23/10/2022

Για τα παρακάτω προβλήματα, καταθέστε ένα αρχείο .R με τις απαντήσεις σας και τα σχετικά σχόλια. Κάθε μέρα καθυστέρησης στην παράδοση της εργασίας έχει penalty μια μονάδα με άριστα το 10.

Προκαταρκτικά/Εξάσκηση
Μια επιπλέον καλή πηγή για εξάσκηση στην dplyr είναι η παρακάτω:
https://r-cubed.rostools.org/ (Section 6).

Όλα όσα θα θέλατε να ξέρετε για τους τύπους δεδομένων στην R:
https://swcarpentry.github.io/r-novice-inflammation/13-supp-data-structures/

[ΜΕΡΟΣ Α] Σας δίνονται τα παρακάτω δεδομένα:

```
data <- tribble(
  ~happy, ~uptempo, ~blues, ~jazz, ~gospel,
  "yes",  "yes",    10,     5,     20,
  "no",   "no",     NA,     12,    15,
  "yes",  "no",     7,      6,     4,
  "no",   "yes",    3,      NA,    NA
)
```

Μετασχηματίστε τα ως εξής:
(α)

```
> data_tidy1
   happy uptempo genre  count
   <chr> <chr>   <chr>  <dbl>
 1 yes   yes     blues     10
 2 no    no      blues     NA
 3 yes   no      blues      7
 4 no    yes     blues      3
 5 yes   yes     jazz       5
 6 no    no      jazz      12
 7 yes   no      jazz       6
 8 no    yes     jazz      NA
 9 yes   yes     gospel    20
10 no    no      gospel    15
11 yes   no      gospel     4
12 no    yes     gospel    NA
```

(β)

```
> data_tidy2
  happy uptempo genre  count
  <chr> <chr>   <chr>  <dbl>
1 yes   yes     blues     10
2 yes   no      blues      7
3 no    yes     blues      3
4 yes   yes     jazz       5
5 no    no      jazz      12
6 yes   no      jazz       6
7 yes   yes     gospel    20
8 no    no      gospel    15
9 yes   no      gospel     4
```

(γ)

```
> data_tidy3
  jazz  happy total
  <lgl> <lgl> <dbl>
1 FALSE FALSE    18
2 FALSE TRUE     41
3 TRUE  FALSE    12
4 TRUE  TRUE     11
```

(δ) Διατυπώστε την κατάλληλη εντολή για να πάρετε το συνολικό πλήθος των λυπημένων τραγουδιών jazz από το data_tidy2. Κάντε το ίδιο για το data_tidy3.

[ΜΕΡΟΣ Β] Στη βιβλιοθήκη tidyverse υπάρχει το dataset who, το οποίο περιέχει στοιχεία του World Health Organization (WHO) σχετικά με τη νόσο της φυματίωσης (TB = tuberculosis).

Προκαταρκτικά.
(α) Φορτώστε τα δεδομένα και μελετήστε τη δομή του πίνακα. Έχει 7240 εγγραφές. Κάθε εγγραφή είναι η καταγραφή δεδομένων της ασθένειας σε κατηγορίες πληθυσμού ανά χώρα και χρονιά.

```
library(tidyverse)
who <- tidyr::who
view(who)
```

(β) Κατεβάστε το αρχείο dict.csv από το https://extranet.who.int/tme/generateCSV.asp?ds=dictionary. Πρόκειται για το λεξικό μεταδεδομένων που περιγράφει τα περιεχόμενα της κάθε μεταβλητής του αρχείου who. Δημιουργούμε έναν νέο πίνακα με τις ονομασίες των στηλών του πίνακα who (labels) και παίρνοντας τη σύζευξη των πινάκων labels και dict μπορούμε να μάθουμε λεπτομέρειες για την κάθε στήλη του πίνακα who.

```
dict_url <- "https://extranet.who.int/tme/generateCSV.asp?ds=dictionary"
if (!file.exists("dict.csv")) download.file(dict_url, "dict.csv")
dict <- read_csv('dict.csv')
view(dict)

labels  <- data.frame(name = colnames(who))
view(labels)

explanations <- semi_join(dict, labels, by=c("variable_name" = "name"))
view(explanations)
```

(1) Μετατρέψτε τον πίνακα ώστε να έχει το ακόλουθο σχήμα: who(country, iso2, iso3, year, notification, cases), όπου οι τιμές της στήλης notification είναι οι new_sp_m014, new_sp_m1524, κλπ. Πρακτικά, όλες οι στήλες που το όνομά τους ξεκινά από new. Προσέξτε ώστε να αγνοήσετε τις τιμές NA. Το αποτέλεσμα θα πρέπει να μοιάζει κάπως έτσι:

```
# A tibble: 76,046 x 6
   country     iso2  iso3   year notification cases
   <chr>       <chr> <chr> <int> <chr>        <int>
 1 Afghanistan AF    AFG    1997 new_sp_m014      0
 2 Afghanistan AF    AFG    1998 new_sp_m014     30
 3 Afghanistan AF    AFG    1999 new_sp_m014      8
 ...
```

(2) Αντικαταστήστε όλες τις εμφανίσεις του string "newrel" στη στήλη notification με το string "new_rel" (ΒΟΗΘΕΙΑ: δείτε τη συνάρτηση str_replace).

(3) "Σπάστε" τα περιεχόμενα της στήλης notification στις στήλες new, type, sex, age. Το αποτέλεσμα θα πρέπει να μοιάζει κάπως έτσι:

```
# A tibble: 76,046 x 9
   country     iso2  iso3   year new   type  sex   age   cases
   <chr>       <chr> <chr> <int> <chr> <chr> <chr> <chr> <int>
 1 Afghanistan AF    AFG    1997 new   sp    m     014       0
 2 Afghanistan AF    AFG    1998 new   sp    m     014      30
 3 Afghanistan AF    AFG    1999 new   sp    m     014       8
 ...
```

(4) "Πετάξτε" τις στήλες new, iso2, iso3. Το τελικό αποτέλεσμα θα πρέπει να μοιάζει κάπως έτσι:

```
# A tibble: 76,046 x 6
   country      year type  sex   age   cases
   <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
 ...
```

Σε περίπτωση που δεν καταφέρετε να δημιουργήσετε τον τελικό πίνακα πάνω στον οποίο θα πρέπει να απαντήσετε τα επόμενα τρία προβλήματα, έχω αποθηκεύσει τον πίνακα σε μορφή R Data Format. Διαβάστε τον με την εντολή:

```
tidy_who <- readRDS(file = "/home/gevan/data/tidy_who.rds")
```

(5) Υπολογίστε για κάθε χώρα το συνολικό αριθμό TB cases. Το αποτέλεσμα θα πρέπει να μοιάζει κάπως έτσι:

```
# A tibble: 219 x 2
   country              count
   <chr>                <int>
 1 Afghanistan         140225
 2 Albania               5335
 3 Algeria             128119
 ...
```

(6) Βρείτε για κάθε χρονιά τη χώρα με το μεγαλύτερο αριθμό smear positive pulmonary TB cases (sp). Το αποτέλεσμα θα πρέπει να μοιάζει κάπως έτσι:

```
# A tibble: 33 x 3
# Groups:   year [33]
    year country cases
   <int> <chr>   <int>
 1  1980 Canada    186
 2  1981 Canada    141
 3  1982 Canada    150
...
```

(7) Μόνο για την Ελλάδα, δώστε ένα πινακάκι με μόνο 3 στήλες: year, f, και m, όπου οι τιμές για τις στήλες m και f είναι το σύνολο των TB cases για γυναίκες και άνδρες αντίστοιχα. Ο πίνακας πρέπει να είναι ταξινομημένος σε φθίνουσα κατάταξη ως προς f+m, δηλαδή το συνολικό πλήθος των TB cases. Το αποτέλεσμα θα πρέπει να μοιάζει κάπως έτσι:

```
# A tibble: 5 x 3
# Groups:   year [5]
   year     f     m
  <int> <int> <int>
1  2006   195   378
2  2008   184   349
3  2007   165   335
...
```