HW #6

1678226 Chanmi Yoo(유찬미)

# Load library   
library(tidyverse)

## -- Attaching packages --------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## √ ggplot2 3.2.1 √ purrr 0.3.3  
## √ tibble 2.1.3 √ dplyr 0.8.3  
## √ tidyr 1.0.0 √ stringr 1.4.0  
## √ readr 1.3.1 √ forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

1. who 자료 중 2006 ~ 2010의 자료만을 뽑아 NewWho에 저장하시오.

NewWho <- who %>%   
 filter(year %in% 2006:2010)

1. NewWho자료를 아래의 기준에 따라 tidy data로 변형하여 NewwhoTidy에 저장하시오.

* type: TB type으로 변수이름의 두 번째 부분 (sp, rel, sn, ep)
* gender: 성별 - ageL, ageU: age group의 하한과 상한을 의미하며 65세 이상 그룹에서는 ageU를 100으로 함
* age: ageL과 ageU의 평균 값

NewwhoTidy <- NewWho %>%   
 # new\_sp\_m014부터 newrel\_f65 변수를 합치기   
 gather(new\_sp\_m014:newrel\_f65, key="key", value="value", na.rm=T) %>%   
 # newrel new\_rel로 변환   
 mutate(key = str\_replace(key, 'newrel', 'new\_rel')) %>%   
 # key열을 new, type, sexage열로 나누기   
 separate(key, into=c('new', 'type', 'sexage'), sep='\_') %>%  
 # iso2, iso3, new열 삭제   
 select(-iso2, -iso3, -new) %>%   
 # sexage열을 sex, age 열로 나누기   
 separate(sexage, into = c('gender', 'age'), sep = 1) %>%   
 # ageL열의 빈 값을 65으로 변경, ageU열의 65값을 100으로 변경   
 mutate(age = str\_replace(age, '014', '0014'),  
 age = str\_replace(age, '65','65100')) %>%   
 # age열을 ageL과 ageU로 나누기   
 separate(age, into = c('ageL', 'ageU'), sep = 2, convert = T) %>%   
 # age열 생성   
 mutate(age = (ageL + ageU)/ 2) %>%   
 # 열 배열 순서 변경   
 select(country, year, type, gender, ageL, age, ageU, value)  
  
head(NewwhoTidy)

## # A tibble: 6 x 8  
## country year type gender ageL age ageU value  
## <chr> <int> <chr> <chr> <int> <dbl> <int> <int>  
## 1 Afghanistan 2006 sp m 0 7 14 193  
## 2 Afghanistan 2007 sp m 0 7 14 186  
## 3 Afghanistan 2008 sp m 0 7 14 187  
## 4 Afghanistan 2009 sp m 0 7 14 200  
## 5 Afghanistan 2010 sp m 0 7 14 197  
## 6 Albania 2006 sp m 0 7 14 5

tail(NewwhoTidy)

## # A tibble: 6 x 8  
## country year type gender ageL age ageU value  
## <chr> <int> <chr> <chr> <int> <dbl> <int> <int>  
## 1 Yemen 2010 ep f 65 82.5 100 0  
## 2 Zimbabwe 2006 ep f 65 82.5 100 92  
## 3 Zimbabwe 2007 ep f 65 82.5 100 9  
## 4 Zimbabwe 2008 ep f 65 82.5 100 104  
## 5 Zimbabwe 2009 ep f 65 82.5 100 138  
## 6 Zimbabwe 2010 ep f 65 82.5 100 146

1. 2에서 만든 NewWhoTidy 자료를 이용하여 나라별, type별, 연도별로 TB 수의 합을 구하고 각 연도별, type별로 구한 TB 수가 가장 큰 나라를 찾으시오.

NewwhoTidy %>%   
 group\_by(country, type, year) %>%   
 summarise(value.count = sum(value)) %>%   
 group\_by(year) %>%   
 mutate(rank = dense\_rank(desc(value.count))) %>%   
 filter(rank==1)

## # A tibble: 5 x 5  
## # Groups: year [5]  
## country type year value.count rank  
## <chr> <chr> <int> <int> <int>  
## 1 India sp 2006 553797 1  
## 2 India sp 2007 592587 1  
## 3 India sp 2008 615492 1  
## 4 India sp 2009 624617 1  
## 5 India sp 2010 630164 1

# [결론] 매년 인도가 가장 많다.   
  
NewwhoTidy %>%   
 group\_by(country, type, year) %>%   
 summarise(value.count = sum(value)) %>%   
 group\_by(type) %>%   
 mutate(rank = dense\_rank(desc(value.count))) %>%   
 filter(rank==1)

## # A tibble: 3 x 5  
## # Groups: type [3]  
## country type year value.count rank  
## <chr> <chr> <int> <int> <int>  
## 1 China sn 2009 435325 1  
## 2 India ep 2007 206840 1  
## 3 India sp 2010 630164 1

# [결론] sn은 중국, ep와 sp는 인도가 가장 많다.