lec 1112

박효선 1585063

2018년 12월 3일

##############  
# 11 월 12 일  
##############  
  
  
# 1980 년부터 2013 년까지의 WHO 의 TB 원자료.  
# – country : 나라이름  
# – iso2,iso2 : 나라이름에 대한 code  
# – new\_sp\_m014 - new\_rel\_f65: 그룹별 새로운 결핵(TB:Tuberculosis)환자수  
# – rel: relapse, sn:negative pulmonary smear, sp:extrapulmonary  
# – f:female, m:male  
# – 014:0~14 세, 1524:15~24 세, …, 65: 65 세이상  
  
who # tidy 아님! # 7,240 x 60

## # A tibble: 7,240 x 60  
## country iso2 iso3 year new\_sp\_m014 new\_sp\_m1524 new\_sp\_m2534  
## <chr> <chr> <chr> <int> <int> <int> <int>  
## 1 Afghanistan AF AFG 1980 NA NA NA  
## 2 Afghanistan AF AFG 1981 NA NA NA  
## 3 Afghanistan AF AFG 1982 NA NA NA  
## 4 Afghanistan AF AFG 1983 NA NA NA  
## 5 Afghanistan AF AFG 1984 NA NA NA  
## 6 Afghanistan AF AFG 1985 NA NA NA  
## 7 Afghanistan AF AFG 1986 NA NA NA  
## 8 Afghanistan AF AFG 1987 NA NA NA  
## 9 Afghanistan AF AFG 1988 NA NA NA  
## 10 Afghanistan AF AFG 1989 NA NA NA  
## # ... with 7,230 more rows, and 53 more variables: new\_sp\_m3544 <int>,  
## # 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>

tail(who)

## # A tibble: 6 x 60  
## country iso2 iso3 year new\_sp\_m014 new\_sp\_m1524 new\_sp\_m2534  
## <chr> <chr> <chr> <int> <int> <int> <int>  
## 1 Zimbabwe ZW ZWE 2008 127 614 0  
## 2 Zimbabwe ZW ZWE 2009 125 578 NA  
## 3 Zimbabwe ZW ZWE 2010 150 710 2208  
## 4 Zimbabwe ZW ZWE 2011 152 784 2467  
## 5 Zimbabwe ZW ZWE 2012 120 783 2421  
## 6 Zimbabwe ZW ZWE 2013 NA NA NA  
## # ... with 53 more variables: new\_sp\_m3544 <int>, 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>

# tidy 자료를 위해 필요한 변수들.  
# 1. country  
# 2. year  
# 3. ? ==> 일단 cases 로 new\_sp\_m014 부터 newrel\_f65 변수를 합치기.  
  
who1 <- who %>% gather(new\_sp\_m014:newrel\_f65, key = "KEY", value = "cases")  
who1 # 405,440 x 6

## # A tibble: 405,440 x 6  
## country iso2 iso3 year KEY cases  
## <chr> <chr> <chr> <int> <chr> <int>  
## 1 Afghanistan AF AFG 1980 new\_sp\_m014 NA  
## 2 Afghanistan AF AFG 1981 new\_sp\_m014 NA  
## 3 Afghanistan AF AFG 1982 new\_sp\_m014 NA  
## 4 Afghanistan AF AFG 1983 new\_sp\_m014 NA  
## 5 Afghanistan AF AFG 1984 new\_sp\_m014 NA  
## 6 Afghanistan AF AFG 1985 new\_sp\_m014 NA  
## 7 Afghanistan AF AFG 1986 new\_sp\_m014 NA  
## 8 Afghanistan AF AFG 1987 new\_sp\_m014 NA  
## 9 Afghanistan AF AFG 1988 new\_sp\_m014 NA  
## 10 Afghanistan AF AFG 1989 new\_sp\_m014 NA  
## # ... with 405,430 more rows

who1 <- who %>% gather(new\_sp\_m014:newrel\_f65, key="KEY", value="cases", na.rm=TRUE)  
who1 # 76,046 x 6 # NA 제거

## # 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 1998 new\_sp\_m014 30  
## 3 Afghanistan AF AFG 1999 new\_sp\_m014 8  
## 4 Afghanistan AF AFG 2000 new\_sp\_m014 52  
## 5 Afghanistan AF AFG 2001 new\_sp\_m014 129  
## 6 Afghanistan AF AFG 2002 new\_sp\_m014 90  
## 7 Afghanistan AF AFG 2003 new\_sp\_m014 127  
## 8 Afghanistan AF AFG 2004 new\_sp\_m014 139  
## 9 Afghanistan AF AFG 2005 new\_sp\_m014 151  
## 10 Afghanistan AF AFG 2006 new\_sp\_m014 193  
## # ... with 76,036 more rows

# KEY 변수 살펴보기  
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

# key 값의 의미  
# 1. 첫부분 “new”  
# 2. 두번째 부분  
# – rel :cases of relapse  
# – ep : cases of extrapulmonary TB  
# – sn : cases of pulmonary TB that could not be diagnosed by a pulmonary smear  
# (smear negative)  
# – sp : cases of pulmonary TB that could be diagnosed be a pulmonary smear  
# (smear positive)  
# 3. 세번째 부분: 환자 성별 males (m) and females (f).  
# 4. 나머지 숫자들: age group  
# – 014 = 0 – 14 years old  
# – 1524 = 15 – 24 years old  
# – 2534 = 25 – 34 years old  
# – 3544 = 35 – 44 years old  
# – 4554 = 45 – 54 years old  
# – 5564 = 55 – 64 years old  
# – 65 = 65 or older  
  
  
# 먼저 newrel 을 new\_rel 로 변환  
who2 <- who1 %>% mutate(KEY = str\_replace(KEY, "newrel", "new\_rel"))

## Warning: package 'bindrcpp' was built under R version 3.5.1

# key 를 searate() 함수를 이용하여 분리  
# 각 값은 하나의 셀에 저장되어야 한다. 하지만 한 셀에 4가지 정보가 있음 문제!  
# seperate 여러 셀로 column 을 나눠준다  
who3 <- who2 %>% separate(KEY, c("new", "type", "sexage"), sep = "\_")  
who3

## # A tibble: 76,046 x 8  
## country iso2 iso3 year new type sexage cases  
## <chr> <chr> <chr> <int> <chr> <chr> <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  
## 4 Afghanistan AF AFG 2000 new sp m014 52  
## 5 Afghanistan AF AFG 2001 new sp m014 129  
## 6 Afghanistan AF AFG 2002 new sp m014 90  
## 7 Afghanistan AF AFG 2003 new sp m014 127  
## 8 Afghanistan AF AFG 2004 new sp m014 139  
## 9 Afghanistan AF AFG 2005 new sp m014 151  
## 10 Afghanistan AF AFG 2006 new sp m014 193  
## # ... with 76,036 more rows

who3 %>% count(new) # new 한가지 타입 뿐이네

## # A tibble: 1 x 2  
## new n  
## <chr> <int>  
## 1 new 76046

who3 %>% count(type) # 4가지 type 이 있다

## # A tibble: 4 x 2  
## type n  
## <chr> <int>  
## 1 ep 14304  
## 2 rel 2580  
## 3 sn 14342  
## 4 sp 44820

who3 %>% count(sexage) # 성별과 나이 붙어져서 문제다!

## # A tibble: 14 x 2  
## sexage n  
## <chr> <int>  
## 1 f014 5436  
## 2 f1524 5421  
## 3 f2534 5419  
## 4 f3544 5423  
## 5 f4554 5422  
## 6 f5564 5412  
## 7 f65 5415  
## 8 m014 5446  
## 9 m1524 5447  
## 10 m2534 5431  
## 11 m3544 5452  
## 12 m4554 5454  
## 13 m5564 5439  
## 14 m65 5429

# 필요 없는 변수 new, iso2, iso3 제거  
who4 <- who3 %>% select(country, year, type, sexage, cases) # 필요한 변수만 골라서 저장  
who4

## # A tibble: 76,046 x 5  
## country year type sexage cases  
## <chr> <int> <chr> <chr> <int>  
## 1 Afghanistan 1997 sp m014 0  
## 2 Afghanistan 1998 sp m014 30  
## 3 Afghanistan 1999 sp m014 8  
## 4 Afghanistan 2000 sp m014 52  
## 5 Afghanistan 2001 sp m014 129  
## 6 Afghanistan 2002 sp m014 90  
## 7 Afghanistan 2003 sp m014 127  
## 8 Afghanistan 2004 sp m014 139  
## 9 Afghanistan 2005 sp m014 151  
## 10 Afghanistan 2006 sp m014 193  
## # ... with 76,036 more rows

# sexage 를 sex 와 age 로 분리  
who5 <- who4 %>% separate(sexage, c("sex", "age"), sep = 1) # sep = 1 앞에서 한자리 분리  
who5

## # 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  
## 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

who4 %>% separate(sexage,c("sex","age"),sep=2) # sep = 2 앞에서 두자리 분리

## # A tibble: 76,046 x 6  
## country year type sex age cases  
## <chr> <int> <chr> <chr> <chr> <int>  
## 1 Afghanistan 1997 sp m0 14 0  
## 2 Afghanistan 1998 sp m0 14 30  
## 3 Afghanistan 1999 sp m0 14 8  
## 4 Afghanistan 2000 sp m0 14 52  
## 5 Afghanistan 2001 sp m0 14 129  
## 6 Afghanistan 2002 sp m0 14 90  
## 7 Afghanistan 2003 sp m0 14 127  
## 8 Afghanistan 2004 sp m0 14 139  
## 9 Afghanistan 2005 sp m0 14 151  
## 10 Afghanistan 2006 sp m0 14 193  
## # ... with 76,036 more rows

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

## # A tibble: 76,046 x 6  
## country year type sex age cases  
## <chr> <int> <chr> <chr> <chr> <int>  
## 1 Afghanistan 1997 sp m01 4 0  
## 2 Afghanistan 1998 sp m01 4 30  
## 3 Afghanistan 1999 sp m01 4 8  
## 4 Afghanistan 2000 sp m01 4 52  
## 5 Afghanistan 2001 sp m01 4 129  
## 6 Afghanistan 2002 sp m01 4 90  
## 7 Afghanistan 2003 sp m01 4 127  
## 8 Afghanistan 2004 sp m01 4 139  
## 9 Afghanistan 2005 sp m01 4 151  
## 10 Afghanistan 2006 sp m01 4 193  
## # ... with 76,036 more rows

who5 %>% count(age)

## # A tibble: 7 x 2  
## age n  
## <chr> <int>  
## 1 014 10882  
## 2 1524 10868  
## 3 2534 10850  
## 4 3544 10875  
## 5 4554 10876  
## 6 5564 10851  
## 7 65 10844

who5 %>% mutate(age = str\_replace(age, "65", "6500")) %>%   
 separate(age, c("ageL","ageU"), sep = -2, convert = TRUE) %>%   
 mutate(ageU = ifelse(ageU == 0 , 80, ageU))

## # A tibble: 76,046 x 7  
## country year type sex ageL ageU cases  
## <chr> <int> <chr> <chr> <int> <dbl> <int>  
## 1 Afghanistan 1997 sp m 0 14 0  
## 2 Afghanistan 1998 sp m 0 14 30  
## 3 Afghanistan 1999 sp m 0 14 8  
## 4 Afghanistan 2000 sp m 0 14 52  
## 5 Afghanistan 2001 sp m 0 14 129  
## 6 Afghanistan 2002 sp m 0 14 90  
## 7 Afghanistan 2003 sp m 0 14 127  
## 8 Afghanistan 2004 sp m 0 14 139  
## 9 Afghanistan 2005 sp m 0 14 151  
## 10 Afghanistan 2006 sp m 0 14 193  
## # ... with 76,036 more rows

# pipe 를 이용하여 한 번에 처리하기  
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

####과제 7  
  
whoTidy <- who %>% gather(code, value, new\_sp\_m014:newrel\_f65, na.rm = TRUE) %>%  
 mutate(code = str\_replace(code, "newrel", "new\_rel")) %>%  
 separate(code, c("new", "type", "genderage")) %>%  
 select(-new, -iso2, -iso3) %>%  
 separate(genderage, c("gender", "age"), sep=1) %>%  
 mutate(age = str\_replace(age, "65", "6500")) %>%  
 separate(age, c("ageL", "ageU"), sep = -2, convert = TRUE) %>%  
 mutate(ageU = ifelse(ageU == 0, 80, ageU),  
 age = (ageL + ageU)/2) %>%  
 select(country, year, type, gender, ageL, age, ageU, value)