Sample WHO Visualization - R

UB Orengo

LIBRARIES USED

```
library(tidyr)
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.4

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v dplyr   1.0.2
## v tibble  3.0.4     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(dplyr)
library(countrycode)

## Warning: package 'countrycode' was built under R version 4.0.4
```

DATASET - WHO DATASET IN TIDYR

```
tidyr::who

## # A tibble: 7,240 x 60
##    country iso2  iso3   year new_sp_m014 new_sp_m1524 new_sp_m2534 new_sp_m3544
##    <chr>   <chr> <chr> <int>       <int>        <int>        <int>        <int>
##  1 Afghan~ AF    AFG    1980          NA           NA           NA           NA
##  2 Afghan~ AF    AFG    1981          NA           NA           NA           NA
##  3 Afghan~ AF    AFG    1982          NA           NA           NA           NA
##  4 Afghan~ AF    AFG    1983          NA           NA           NA           NA
##  5 Afghan~ AF    AFG    1984          NA           NA           NA           NA
##  6 Afghan~ AF    AFG    1985          NA           NA           NA           NA
##  7 Afghan~ AF    AFG    1986          NA           NA           NA           NA
##  8 Afghan~ AF    AFG    1987          NA           NA           NA           NA
##  9 Afghan~ AF    AFG    1988          NA           NA           NA           NA
## 10 Afghan~ AF    AFG    1989          NA           NA           NA           NA
## # ... with 7,230 more rows, and 52 more variables: 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>
```

INITIAL TIDYING OF DATASET – CREDIT TO R FOR DATA SCIENCE BY WICKAM

LINK TO CLEANUP CODE – https://r4ds.had.co.nz/tidy-data.html

```
who5 <-who %>%
  pivot_longer(
    cols = new_sp_m014:newrel_f65,
    names_to = "key",
    values_to = "cases",
    values_drop_na = TRUE
  ) %>%
  mutate(
    key = stringr::str_replace(key, "newrel", "new_rel")
  ) %>%
  separate(key, c("new", "var", "sexage")) %>%
  select(-new, -iso2, -iso3) %>%
  separate(sexage, c("sex", "age"), sep = 1)

head(who5)
```

```
## # A tibble: 6 x 6
##   country      year var   sex   age   cases
##   <chr>       <int> <chr> <chr> <chr> <int>
## 1 Afghanistan  1997 sp    m     014       0
## 2 Afghanistan  1997 sp    m     1524     10
## 3 Afghanistan  1997 sp    m     2534      6
## 4 Afghanistan  1997 sp    m     3544      3
## 5 Afghanistan  1997 sp    m     4554      5
## 6 Afghanistan  1997 sp    m     5564      2
```

Removed Serbia and Serbia & Montenegro from DF, further in the document, I created a df for continent and cbind'd it to the cleaned WHO data, but it was unable to id those countries. Because it represented such a small subset, I removed them

```
who5_clean <- who5[!(who5$country =="Serbia" | who5$country == "Serbia & Montenegro"),]
```

Converted it back to who5 DF

```
who5 <- who5_clean
```

Created continents list using package "countrycode"

```
continents <- countrycode(sourcevar = who5$country,
                          origin = "country.name",
                          destination = "continent")
```

Converted that into a DF

```
continents_df <- data.frame(continent = continents)
str(continents_df)

## 'data.frame':    75626 obs. of  1 variable:
##  $ continent: chr  "Asia" "Asia" "Asia" "Asia" ...
```

Bound that DF with the rest of the dataframe for who5

```
who5_with_continent <- cbind(continents_df,who5)

head(who5_with_continent)

##   continent     country year var sex  age cases
## 1      Asia Afghanistan 1997  sp   m  014     0
## 2      Asia Afghanistan 1997  sp   m 1524    10
## 3      Asia Afghanistan 1997  sp   m 2534     6
## 4      Asia Afghanistan 1997  sp   m 3544     3
## 5      Asia Afghanistan 1997  sp   m 4554     5
## 6      Asia Afghanistan 1997  sp   m 5564     2
```

summarized the data by continent, year, and sex using group_by & summarize from the dplyr package

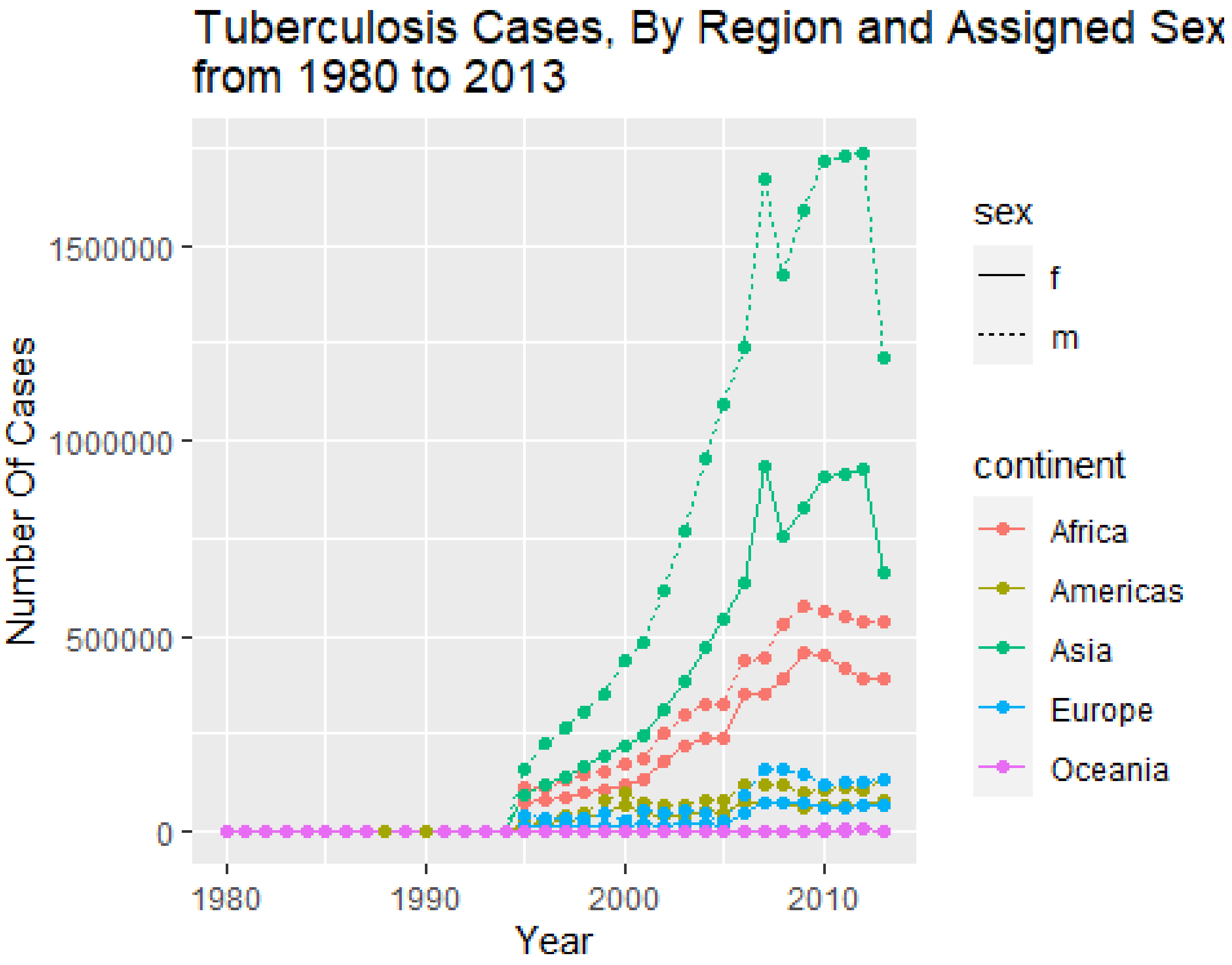
```
who_summ <- who5_with_continent %>%
  group_by(continent, year, sex) %>%
  summarise(number_of_cases = sum(cases))

## `summarise()` regrouping output by 'continent', 'year' (override with `.groups`
argument)

head(who_summ)

## # A tibble: 6 x 4
## # Groups:   continent, year [3]
##   continent  year sex   number_of_cases
##   <chr>     <int> <chr>           <int>
## 1 Africa     1995 f               71394
## 2 Africa     1995 m              109117
## 3 Africa     1996 f               76536
## 4 Africa     1996 m              115154
## 5 Africa     1997 f               85606
## 6 Africa     1997 m              129248
```

Plotted the data, using x axis for year and y axis for number of cases. I color coded it with the continent variable and created two linetypes for sex

```
who_viz <- ggplot(data = who_summ, mapping = aes(x = year, y = number_of_cases, color =
continent)) +
          geom_line(aes(linetype = sex)) +
          geom_point()
```

Printed the plot, added title and cleaned up labels

```
print(who_viz + ggtitle("Tuberculosis Cases, By Region and Assigned Sex At Birth \nfrom
1980 to 2013") +labs(x = "Year", y = "Number Of Cases"))
```


Tuberculosis Cases, By Region and Assigned Sex
from 1980 to 2013
Number Of Cases
1500000
1000000
500000
0
1980
1990
2000
2010
Year
sex
f
m
continent
Africa
Americas
Asia
Europe
Oceania