

Run 4 Course 1 Week 2 Office Hour

Jiaying Wu

08/09/2020

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse
```

```
## v ggplot2 3.3.2    v purrr  0.3.4
## v tibble  3.0.3    v dplyr  1.0.1
## v tidyr   1.1.1    v stringr 1.4.0
## v readr   1.3.1    v forcats 0.5.0
```

```
## -- Conflicts ----- tidyverse_conflicts()
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

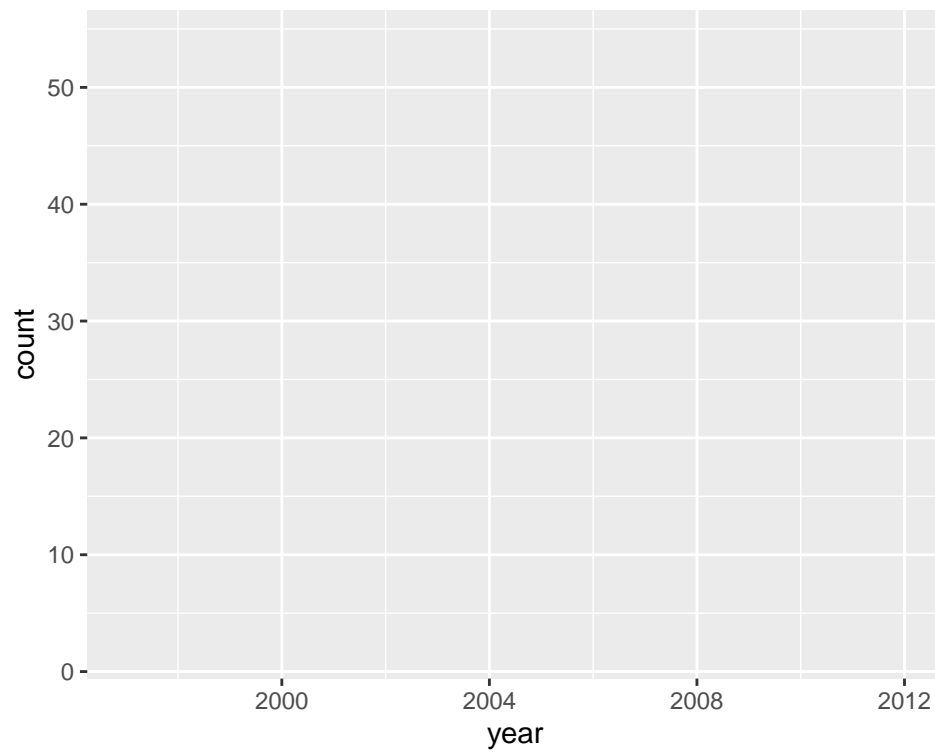
```
# load data and filter the aus without missing value in age group
```

```
tb <- read_rds(here::here("data", "tb_tidy.rds"))
tb_au <- filter(tb, country == "Australia", !is.na(age_group))
tb_au
```

```
## # A tibble: 192 x 6
##   country iso3  year count sex  age_group
##   <chr>   <chr> <dbl> <dbl> <fct> <fct>
## 1 Australia AUS   1997     8 M    15-24
## 2 Australia AUS   1998    11 M    15-24
## 3 Australia AUS   1999    13 M    15-24
## 4 Australia AUS   2000    16 M    15-24
## 5 Australia AUS   2001    23 M    15-24
## 6 Australia AUS   2002    15 M    15-24
## 7 Australia AUS   2003    14 M    15-24
## 8 Australia AUS   2004    18 M    15-24
## 9 Australia AUS   2005    32 M    15-24
## 10 Australia AUS   2006    33 M    15-24
## # ... with 182 more rows
```

```
# blank plot
```

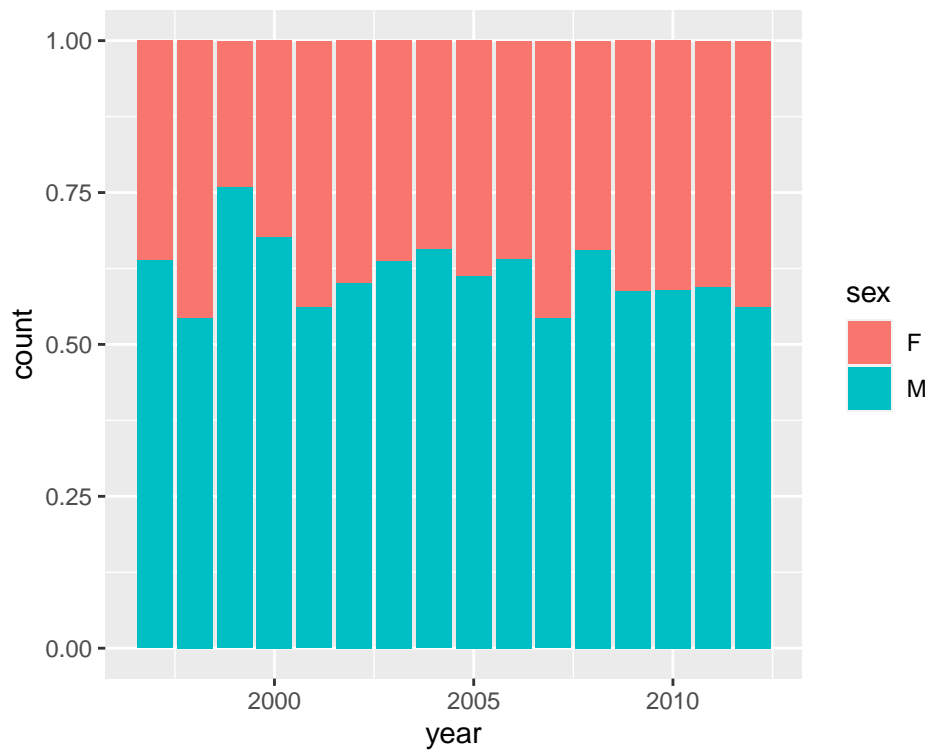
```
p1 <- ggplot(tb_au, aes(x = year, y = count, fill = sex))
p1
```



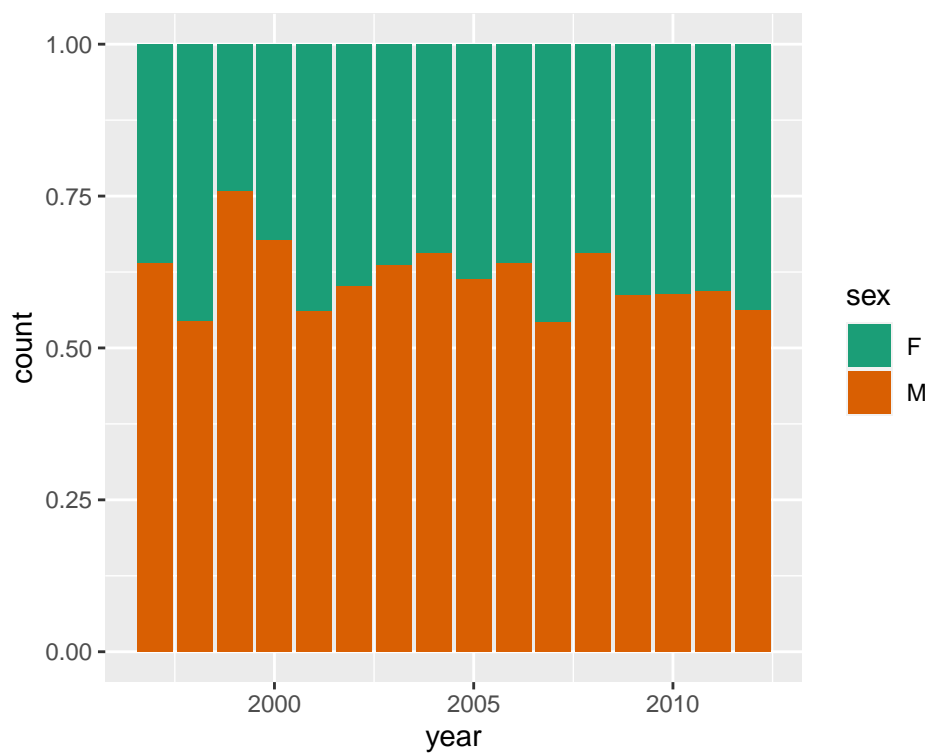
Bar chart

100% Bar Chart

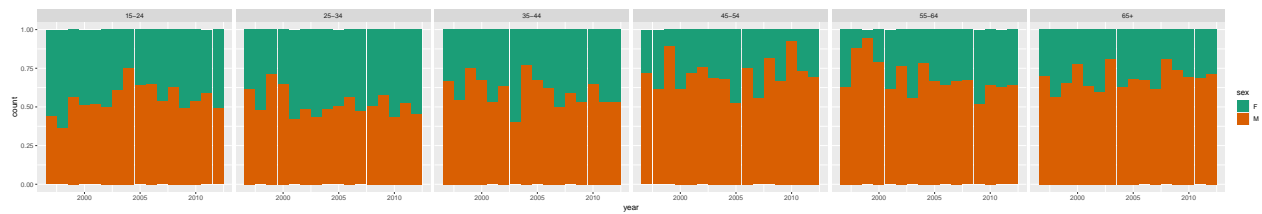
```
# 100% bar chart  
p2 <- p1 + geom_bar(stat = "identity", position = "fill")  
p2
```



```
# change color which friendly for colour blindness
p3 <- p2 + scale_fill_brewer(palette = "Dark2")
p3
```



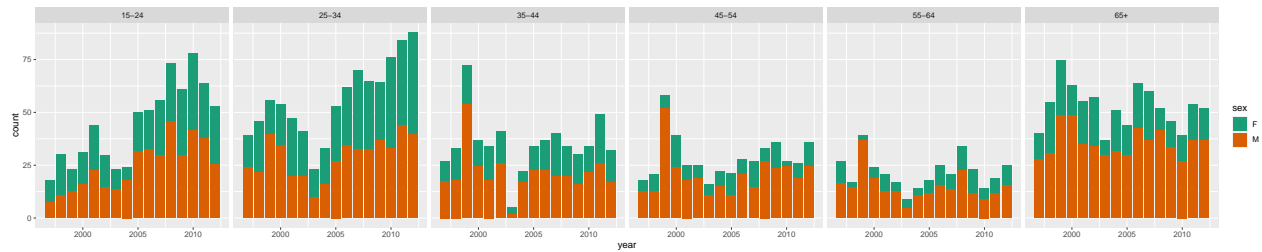
```
p4 <- p3 + facet_grid(~age_group)
p4
```



Stack bar chart

```
stack_bar <- ggplot(tb_au, aes(x = year, y = count, fill = sex)) +
  geom_bar(stat = "identity", position = "stack") +
  facet_grid(~ age_group) +
  scale_fill_brewer(palette="Dark2")
```

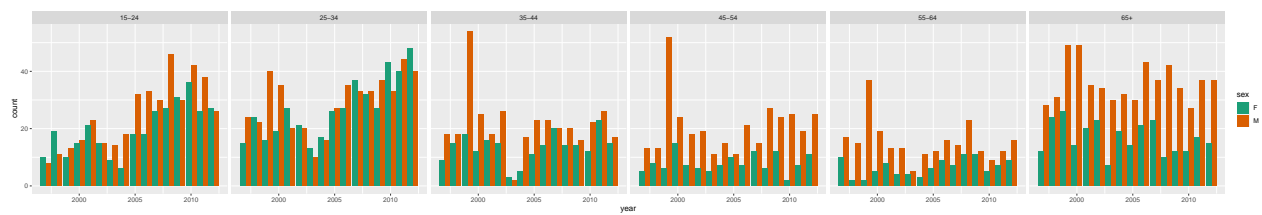
stack_bar



Side by side bar chart (dodge)

```
sbs_bar <- ggplot(tb_au, aes(x = year, y = count, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_grid(~ age_group) +
  scale_fill_brewer(palette="Dark2")
```

sbs_bar



Separated bar chart

```
sep_bar <- ggplot(tb_au, aes(x = year, y = count, fill = sex)) +
  geom_bar(stat = "identity") +
  facet_grid(sex ~ age_group) +
  scale_fill_brewer(palette="Dark2")
```

sep_bar

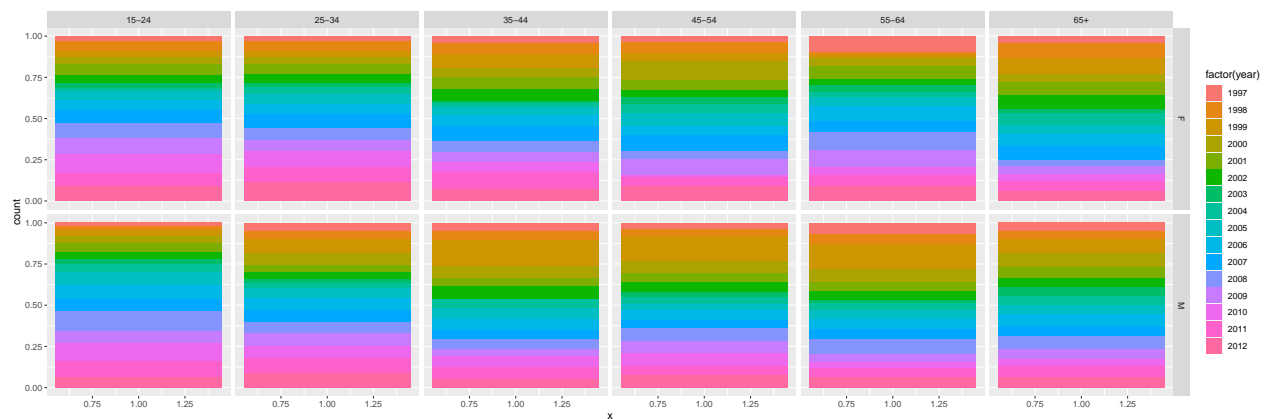


Rainbow, rose and pie charts

Rainbow chart

```
rainbow <- ggplot(tb_au, aes(x = 1, y = count, fill = factor(year))) +
  geom_bar(stat = "identity", position="fill") +
  facet_grid(sex ~ age_group)
```

rainbow



Rose chart

```

rose <- ggplot(tb_au, aes(x = year, y = count, fill = sex)) +
  geom_bar(stat = "identity") +
  facet_grid(sex ~ age_group) +
  scale_fill_brewer(palette="Dark2") +
  # change to rose chart
  coord_polar()

```

rose



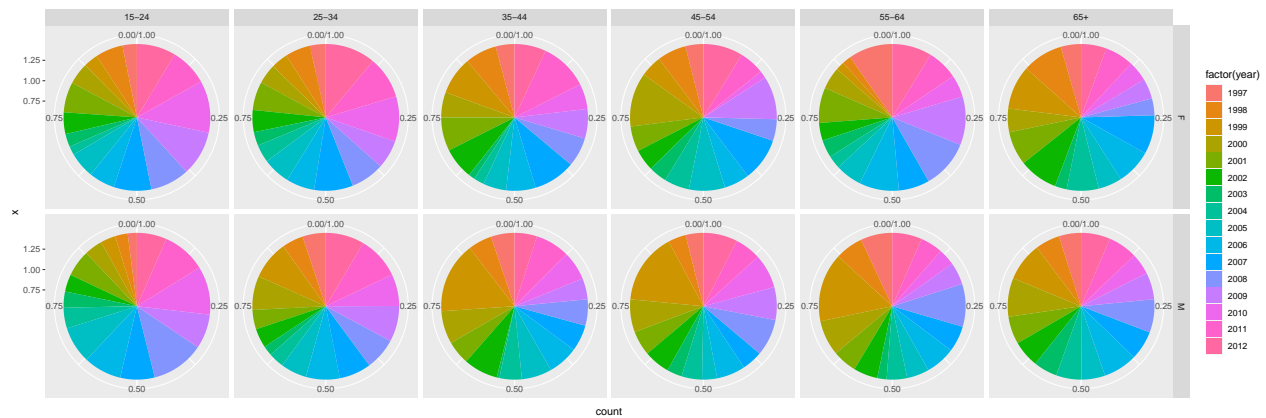
Rainbow pie chart

```

rainbow_pie <- rainbow +
  # change to pie chart
  coord_polar(theta = "y")

```

rainbow_pie



Tidy data

What is tidy data?

- Each variable is in a column.

- Each observation is a row.
- Each value is a cell.

gather function

```
tb_smaller <- tibble(year = c(2016, 2017, 2018),
                     male = c(10, 20, 30),
                     female = c(5, 15, 12))
tb_smaller
```

```
## # A tibble: 3 x 3
##   year  male female
##   <dbl> <dbl> <dbl>
## 1  2016     10      5
## 2  2017     20     15
## 3  2018     30     12
```

```
tb_smaller_long <- gather(tb_smaller,
                          # key is the value in the column names
                          key = "sex",
                          # value is the value in the cell
                          value = "count",
                          # the variable take into account in the tb_smaller data
                          male, female)
tb_smaller_long
```

```
## # A tibble: 6 x 3
##   year sex    count
##   <dbl> <chr>  <dbl>
## 1  2016 male     10
## 2  2017 male     20
## 3  2018 male     30
## 4  2016 female     5
## 5  2017 female    15
## 6  2018 female    12
```

tidy format

```
tb
```

```
## # A tibble: 47,866 x 6
##   country    iso3  year count sex  age_group
##   <chr>      <chr> <dbl> <dbl> <fct> <fct>
## 1 Afghanistan AFG   1997    10 M    15-24
## 2 Afghanistan AFG   1998   129 M    15-24
## 3 Afghanistan AFG   1999    55 M    15-24
## 4 Afghanistan AFG   2000   228 M    15-24
## 5 Afghanistan AFG   2001   379 M    15-24
```

```
## 6 Afghanistan AFG 2002 476 M 15-24
## 7 Afghanistan AFG 2003 511 M 15-24
## 8 Afghanistan AFG 2004 537 M 15-24
## 9 Afghanistan AFG 2005 606 M 15-24
## 10 Afghanistan AFG 2006 837 M 15-24
## # ... with 47,856 more rows
```

messey format

```
tb_messy <- read_csv("data/TB_notifications.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   country = col_character(),
##   iso3 = col_character()
## )

## See spec(...) for full column specifications.
```

```
tb_messy
```

```
## # A tibble: 7,891 x 23
##   country iso3 year new_sp_m04 new_sp_m514 new_sp_m014 new_sp_m1524
##   <chr>   <chr> <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 Afghan~ AFG 1980         NA         NA         NA         NA
## 2 Afghan~ AFG 1981         NA         NA         NA         NA
## 3 Afghan~ AFG 1982         NA         NA         NA         NA
## 4 Afghan~ AFG 1983         NA         NA         NA         NA
## 5 Afghan~ AFG 1984         NA         NA         NA         NA
## 6 Afghan~ AFG 1985         NA         NA         NA         NA
## 7 Afghan~ AFG 1986         NA         NA         NA         NA
## 8 Afghan~ AFG 1987         NA         NA         NA         NA
## 9 Afghan~ AFG 1988         NA         NA         NA         NA
## 10 Afghan~ AFG 1989         NA         NA         NA         NA
## # ... with 7,881 more rows, and 16 more variables: new_sp_m2534 <dbl>,
## #   new_sp_m3544 <dbl>, new_sp_m4554 <dbl>, new_sp_m5564 <dbl>,
## #   new_sp_m65 <dbl>, new_sp_mu <dbl>, new_sp_f04 <dbl>, new_sp_f514 <dbl>,
## #   new_sp_f014 <dbl>, new_sp_f1524 <dbl>, new_sp_f2534 <dbl>,
## #   new_sp_f3544 <dbl>, new_sp_f4554 <dbl>, new_sp_f5564 <dbl>,
## #   new_sp_f65 <dbl>, new_sp_fu <dbl>
```

Convert data in from wider format to longer format

```
tb_long <- gather(tb_messy, key = "sex_agegroup", value = "count", -country, -year, -iso3)
tb_long
```

```
## # A tibble: 157,820 x 5
```



```
##   country      iso3   year sex_agegroup count
##   <chr>        <chr> <dbl> <chr>         <dbl>
## 1 Afghanistan AFG    1980 new_sp_m04      NA
## 2 Afghanistan AFG    1981 new_sp_m04      NA
## 3 Afghanistan AFG    1982 new_sp_m04      NA
## 4 Afghanistan AFG    1983 new_sp_m04      NA
## 5 Afghanistan AFG    1984 new_sp_m04      NA
## 6 Afghanistan AFG    1985 new_sp_m04      NA
## 7 Afghanistan AFG    1986 new_sp_m04      NA
## 8 Afghanistan AFG    1987 new_sp_m04      NA
## 9 Afghanistan AFG    1988 new_sp_m04      NA
## 10 Afghanistan AFG    1989 new_sp_m04      NA
## # ... with 157,810 more rows
```

Separate “sex_agegroup”

```
tb_long2 <- separate(tb_long,
  col = "sex_agegroup",
  into = c("type", "sex_agegroup"),
  # split at the underscore that is followed by the characters "m" or "f"
  sep = "_(?=[mf])"
)
tb_long2
```

```
## # A tibble: 157,820 x 6
##   country      iso3   year type  sex_agegroup count
##   <chr>        <chr> <dbl> <chr> <chr>         <dbl>
## 1 Afghanistan AFG    1980 new_sp m04      NA
## 2 Afghanistan AFG    1981 new_sp m04      NA
## 3 Afghanistan AFG    1982 new_sp m04      NA
## 4 Afghanistan AFG    1983 new_sp m04      NA
## 5 Afghanistan AFG    1984 new_sp m04      NA
## 6 Afghanistan AFG    1985 new_sp m04      NA
## 7 Afghanistan AFG    1986 new_sp m04      NA
## 8 Afghanistan AFG    1987 new_sp m04      NA
## 9 Afghanistan AFG    1988 new_sp m04      NA
## 10 Afghanistan AFG    1989 new_sp m04      NA
## # ... with 157,810 more rows
```

Extract the “age_group”

```
tb_long3 <- extract(tb_long2,
  col = "sex_agegroup",
  into = c("sex", "age_group"),
  # the first says match characters "m" or "f"
  # the second says match anything that comes after.
  regex = "([mf])(.*)")
tb_long3
```

```
## # A tibble: 157,820 x 7
```

```
##   country      iso3   year type    sex  age_group count
##   <chr>        <chr> <dbl> <chr>  <chr> <chr>      <dbl>
##  1 Afghanistan AFG    1980 new_sp m      04         NA
##  2 Afghanistan AFG    1981 new_sp m      04         NA
##  3 Afghanistan AFG    1982 new_sp m      04         NA
##  4 Afghanistan AFG    1983 new_sp m      04         NA
##  5 Afghanistan AFG    1984 new_sp m      04         NA
##  6 Afghanistan AFG    1985 new_sp m      04         NA
##  7 Afghanistan AFG    1986 new_sp m      04         NA
##  8 Afghanistan AFG    1987 new_sp m      04         NA
##  9 Afghanistan AFG    1988 new_sp m      04         NA
## 10 Afghanistan AFG    1989 new_sp m      04         NA
## # ... with 157,810 more rows
```

Wrangling data

wrangling verbs:

- filter (To pick cases/rows)
- select (Select columns, and ignores all others)
- mutate (Add columns)
- summarise (Calculate a summary)
- group_by (Group data)
- arrange (Sort rows)

Filter data

```
tb_long
```

```
## # A tibble: 157,820 x 5
##   country      iso3   year sex_agegroup count
##   <chr>        <chr> <dbl> <chr>      <dbl>
##  1 Afghanistan AFG    1980 new_sp_m04    NA
##  2 Afghanistan AFG    1981 new_sp_m04    NA
##  3 Afghanistan AFG    1982 new_sp_m04    NA
##  4 Afghanistan AFG    1983 new_sp_m04    NA
##  5 Afghanistan AFG    1984 new_sp_m04    NA
##  6 Afghanistan AFG    1985 new_sp_m04    NA
##  7 Afghanistan AFG    1986 new_sp_m04    NA
##  8 Afghanistan AFG    1987 new_sp_m04    NA
##  9 Afghanistan AFG    1988 new_sp_m04    NA
## 10 Afghanistan AFG    1989 new_sp_m04    NA
## # ... with 157,810 more rows
```

```
# filter country equal to Australia
tb_au <- filter(tb_long, country == "Australia")
tb_au
```

```
## # A tibble: 740 x 5
##   country iso3   year sex_agegroup count
##   <chr>   <chr> <dbl> <chr>      <dbl>
## 1 Australia AUS    1980 new_sp_m04    NA
## 2 Australia AUS    1981 new_sp_m04    NA
## 3 Australia AUS    1982 new_sp_m04    NA
## 4 Australia AUS    1983 new_sp_m04    NA
## 5 Australia AUS    1984 new_sp_m04    NA
## 6 Australia AUS    1985 new_sp_m04    NA
## 7 Australia AUS    1986 new_sp_m04    NA
## 8 Australia AUS    1987 new_sp_m04    NA
## 9 Australia AUS    1988 new_sp_m04    NA
## 10 Australia AUS    1989 new_sp_m04    NA
## # ... with 730 more rows
```

```
# filter count larger than 10
filter(tb_long, count >= 10)
```

```
## # A tibble: 31,507 x 5
##   country iso3   year sex_agegroup count
##   <chr>   <chr> <dbl> <chr>      <dbl>
## 1 Angola   AGO    2011 new_sp_m04    108
## 2 Angola   AGO    2012 new_sp_m04     58
## 3 Argentina ARG    2006 new_sp_m04     19
## 4 Argentina ARG    2007 new_sp_m04     14
## 5 Argentina ARG    2008 new_sp_m04     11
## 6 Argentina ARG    2010 new_sp_m04     13
## 7 Argentina ARG    2011 new_sp_m04     50
## 8 Botswana BWA    2009 new_sp_m04     12
## 9 Botswana BWA    2010 new_sp_m04     11
## 10 Botswana BWA    2011 new_sp_m04     14
## # ... with 31,497 more rows
```

```
filter(tb_long, iso3 %in% c("AUS", "NZL", "IDN"))
```

```
## # A tibble: 2,220 x 5
##   country iso3   year sex_agegroup count
##   <chr>   <chr> <dbl> <chr>      <dbl>
## 1 Australia AUS    1980 new_sp_m04    NA
## 2 Australia AUS    1981 new_sp_m04    NA
## 3 Australia AUS    1982 new_sp_m04    NA
## 4 Australia AUS    1983 new_sp_m04    NA
## 5 Australia AUS    1984 new_sp_m04    NA
## 6 Australia AUS    1985 new_sp_m04    NA
## 7 Australia AUS    1986 new_sp_m04    NA
## 8 Australia AUS    1987 new_sp_m04    NA
## 9 Australia AUS    1988 new_sp_m04    NA
## 10 Australia AUS    1989 new_sp_m04    NA
## # ... with 2,210 more rows
```

```
filter(tb_long, !is.na(count))
```

```
## # A tibble: 50,974 x 5
##   country    iso3   year sex_agegroup count
##   <chr>      <chr> <dbl> <chr>      <dbl>
## 1 Afghanistan AFG    2010 new_sp_m04      4
## 2 Afghanistan AFG    2011 new_sp_m04      2
## 3 Afghanistan AFG    2012 new_sp_m04      0
## 4 Albania      ALB    2005 new_sp_m04      0
## 5 Albania      ALB    2006 new_sp_m04      1
## 6 Albania      ALB    2007 new_sp_m04      0
## 7 Albania      ALB    2008 new_sp_m04      1
## 8 Albania      ALB    2009 new_sp_m04      0
## 9 Albania      ALB    2010 new_sp_m04      0
## 10 Albania     ALB    2011 new_sp_m04      0
## # ... with 50,964 more rows
```

```
# filter base on two condition
filter(tb_long, !is.na(count) & country == "India")
```

```
## # A tibble: 252 x 5
##   country iso3   year sex_agegroup count
##   <chr>    <chr> <dbl> <chr>      <dbl>
## 1 India   IND    1995 new_sp_m014     16
## 2 India   IND    1996 new_sp_m014     47
## 3 India   IND    1997 new_sp_m014     50
## 4 India   IND    1998 new_sp_m014     84
## 5 India   IND    1999 new_sp_m014    327
## 6 India   IND    2000 new_sp_m014   1588
## 7 India   IND    2001 new_sp_m014   1063
## 8 India   IND    2002 new_sp_m014   2551
## 9 India   IND    2003 new_sp_m014   2411
## 10 India  IND    2004 new_sp_m014   3018
## # ... with 242 more rows
```

```
# Another way to filter base on two condition
filter(tb_long, !is.na(count), country == "India")
```

```
## # A tibble: 252 x 5
##   country iso3   year sex_agegroup count
##   <chr>    <chr> <dbl> <chr>      <dbl>
## 1 India   IND    1995 new_sp_m014     16
## 2 India   IND    1996 new_sp_m014     47
## 3 India   IND    1997 new_sp_m014     50
## 4 India   IND    1998 new_sp_m014     84
## 5 India   IND    1999 new_sp_m014    327
## 6 India   IND    2000 new_sp_m014   1588
## 7 India   IND    2001 new_sp_m014   1063
## 8 India   IND    2002 new_sp_m014   2551
## 9 India   IND    2003 new_sp_m014   2411
## 10 India  IND    2004 new_sp_m014   3018
## # ... with 242 more rows
```

```
# filter union of data not missing in count or country equal to India
filter(tb_long, !is.na(count) | country == "India")
```

```
## # A tibble: 51,462 x 5
##   country      iso3  year sex_agegroup count
##   <chr>        <chr> <dbl> <chr>      <dbl>
## 1 Afghanistan AFG    2010 new_sp_m04      4
## 2 Afghanistan AFG    2011 new_sp_m04      2
## 3 Afghanistan AFG    2012 new_sp_m04      0
## 4 Albania      ALB    2005 new_sp_m04      0
## 5 Albania      ALB    2006 new_sp_m04      1
## 6 Albania      ALB    2007 new_sp_m04      0
## 7 Albania      ALB    2008 new_sp_m04      1
## 8 Albania      ALB    2009 new_sp_m04      0
## 9 Albania      ALB    2010 new_sp_m04      0
## 10 Albania     ALB    2011 new_sp_m04      0
## # ... with 51,452 more rows
```