Week5CH12

Nitin

June 8, 2019

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.3

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

## v ggplot2 3.2.0 v purrr 0.2.5   
## v tibble 2.1.3 v dplyr 0.8.0.1  
## v tidyr 0.8.1 v stringr 1.3.1   
## v readr 1.1.1 v forcats 0.3.0

## Warning: package 'ggplot2' was built under R version 3.5.3

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

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

library(tinytex)

## Warning: package 'tinytex' was built under R version 3.5.3

table1

## # A tibble: 6 x 4  
## country year cases population  
## <chr> <int> <int> <int>  
## 1 Afghanistan 1999 745 19987071  
## 2 Afghanistan 2000 2666 20595360  
## 3 Brazil 1999 37737 172006362  
## 4 Brazil 2000 80488 174504898  
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583

table2

## # A tibble: 12 x 4  
## country year type count  
## <chr> <int> <chr> <int>  
## 1 Afghanistan 1999 cases 745  
## 2 Afghanistan 1999 population 19987071  
## 3 Afghanistan 2000 cases 2666  
## 4 Afghanistan 2000 population 20595360  
## 5 Brazil 1999 cases 37737  
## 6 Brazil 1999 population 172006362  
## 7 Brazil 2000 cases 80488  
## 8 Brazil 2000 population 174504898  
## 9 China 1999 cases 212258  
## 10 China 1999 population 1272915272  
## 11 China 2000 cases 213766  
## 12 China 2000 population 1280428583

table3

## # A tibble: 6 x 3  
## country year rate   
## \* <chr> <int> <chr>   
## 1 Afghanistan 1999 745/19987071   
## 2 Afghanistan 2000 2666/20595360   
## 3 Brazil 1999 37737/172006362   
## 4 Brazil 2000 80488/174504898   
## 5 China 1999 212258/1272915272  
## 6 China 2000 213766/1280428583

table4a

## # A tibble: 3 x 3  
## country `1999` `2000`  
## \* <chr> <int> <int>  
## 1 Afghanistan 745 2666  
## 2 Brazil 37737 80488  
## 3 China 212258 213766

table4b

## # A tibble: 3 x 3  
## country `1999` `2000`  
## \* <chr> <int> <int>  
## 1 Afghanistan 19987071 20595360  
## 2 Brazil 172006362 174504898  
## 3 China 1272915272 1280428583

# Compute rate per 10,000

table1 %>%   
 mutate(rate = cases / population \* 10000)

## # A tibble: 6 x 5  
## country year cases population rate  
## <chr> <int> <int> <int> <dbl>  
## 1 Afghanistan 1999 745 19987071 0.373  
## 2 Afghanistan 2000 2666 20595360 1.29   
## 3 Brazil 1999 37737 172006362 2.19   
## 4 Brazil 2000 80488 174504898 4.61   
## 5 China 1999 212258 1272915272 1.67   
## 6 China 2000 213766 1280428583 1.67

# Compute cases per year

table1 %>%   
 count(year, wt = cases)

## # A tibble: 2 x 2  
## year n  
## <int> <int>  
## 1 1999 250740  
## 2 2000 296920

# Visualise changes over time

library(ggplot2)  
ggplot(table1, aes(year, cases)) +   
 geom\_line(aes(group = country), colour = "grey50") +   
 geom\_point(aes(colour = country))

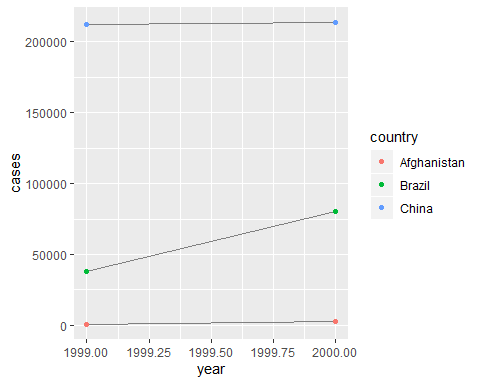


table4a

## # A tibble: 3 x 3  
## country `1999` `2000`  
## \* <chr> <int> <int>  
## 1 Afghanistan 745 2666  
## 2 Brazil 37737 80488  
## 3 China 212258 213766

table4a %>%   
 gather(`1999`, `2000`, key = "year", value = "cases")

## # A tibble: 6 x 3  
## country year cases  
## <chr> <chr> <int>  
## 1 Afghanistan 1999 745  
## 2 Brazil 1999 37737  
## 3 China 1999 212258  
## 4 Afghanistan 2000 2666  
## 5 Brazil 2000 80488  
## 6 China 2000 213766

table4b %>%   
 gather(`1999`, `2000`, key = "year", value = "population")

## # A tibble: 6 x 3  
## country year population  
## <chr> <chr> <int>  
## 1 Afghanistan 1999 19987071  
## 2 Brazil 1999 172006362  
## 3 China 1999 1272915272  
## 4 Afghanistan 2000 20595360  
## 5 Brazil 2000 174504898  
## 6 China 2000 1280428583

tidy4a <- table4a %>%   
 gather(`1999`, `2000`, key = "year", value = "cases")  
tidy4b <- table4b %>%   
 gather(`1999`, `2000`, key = "year", value = "population")  
left\_join(tidy4a, tidy4b)

## Joining, by = c("country", "year")

## # A tibble: 6 x 4  
## country year cases population  
## <chr> <chr> <int> <int>  
## 1 Afghanistan 1999 745 19987071  
## 2 Brazil 1999 37737 172006362  
## 3 China 1999 212258 1272915272  
## 4 Afghanistan 2000 2666 20595360  
## 5 Brazil 2000 80488 174504898  
## 6 China 2000 213766 1280428583

table2

## # A tibble: 12 x 4  
## country year type count  
## <chr> <int> <chr> <int>  
## 1 Afghanistan 1999 cases 745  
## 2 Afghanistan 1999 population 19987071  
## 3 Afghanistan 2000 cases 2666  
## 4 Afghanistan 2000 population 20595360  
## 5 Brazil 1999 cases 37737  
## 6 Brazil 1999 population 172006362  
## 7 Brazil 2000 cases 80488  
## 8 Brazil 2000 population 174504898  
## 9 China 1999 cases 212258  
## 10 China 1999 population 1272915272  
## 11 China 2000 cases 213766  
## 12 China 2000 population 1280428583

table2 %>%  
 spread(key = type, value = count)

## # A tibble: 6 x 4  
## country year cases population  
## <chr> <int> <int> <int>  
## 1 Afghanistan 1999 745 19987071  
## 2 Afghanistan 2000 2666 20595360  
## 3 Brazil 1999 37737 172006362  
## 4 Brazil 2000 80488 174504898  
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583

stocks <- tibble(  
 year = c(2015, 2015, 2016, 2016),  
 half = c( 1, 2, 1, 2),  
 return = c(1.88, 0.59, 0.92, 0.17)  
)  
stocks %>%   
 spread(year, return) %>%   
 gather("year", "return", `2015`:`2016`)

## # A tibble: 4 x 3  
## half year return  
## <dbl> <chr> <dbl>  
## 1 1 2015 1.88  
## 2 2 2015 0.59  
## 3 1 2016 0.92  
## 4 2 2016 0.17

#table4a %>%   
# gather(1999, 2000, key = "year", value = "cases")

people <- tribble(  
 ~name, ~key, ~value,  
 #-----------------|--------|------  
 "Phillip Woods", "age", 45,  
 "Phillip Woods", "height", 186,  
 "Phillip Woods", "age", 50,  
 "Jessica Cordero", "age", 37,  
 "Jessica Cordero", "height", 156  
)

preg <- tribble(  
 ~pregnant, ~male, ~female,  
 "yes", NA, 10,  
 "no", 20, 12  
)

table3

## # A tibble: 6 x 3  
## country year rate   
## \* <chr> <int> <chr>   
## 1 Afghanistan 1999 745/19987071   
## 2 Afghanistan 2000 2666/20595360   
## 3 Brazil 1999 37737/172006362   
## 4 Brazil 2000 80488/174504898   
## 5 China 1999 212258/1272915272  
## 6 China 2000 213766/1280428583

table3 %>%   
 separate(rate, into = c("cases", "population"))

## # A tibble: 6 x 4  
## country year cases population  
## <chr> <int> <chr> <chr>   
## 1 Afghanistan 1999 745 19987071   
## 2 Afghanistan 2000 2666 20595360   
## 3 Brazil 1999 37737 172006362   
## 4 Brazil 2000 80488 174504898   
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583

table3 %>%   
 separate(rate, into = c("cases", "population"), sep = "/")

## # A tibble: 6 x 4  
## country year cases population  
## <chr> <int> <chr> <chr>   
## 1 Afghanistan 1999 745 19987071   
## 2 Afghanistan 2000 2666 20595360   
## 3 Brazil 1999 37737 172006362   
## 4 Brazil 2000 80488 174504898   
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583

table3 %>%   
 separate(rate, into = c("cases", "population"), convert = TRUE)

## # A tibble: 6 x 4  
## country year cases population  
## <chr> <int> <int> <int>  
## 1 Afghanistan 1999 745 19987071  
## 2 Afghanistan 2000 2666 20595360  
## 3 Brazil 1999 37737 172006362  
## 4 Brazil 2000 80488 174504898  
## 5 China 1999 212258 1272915272  
## 6 China 2000 213766 1280428583

table3 %>%   
 separate(year, into = c("century", "year"), sep = 2)

## # A tibble: 6 x 4  
## country century year rate   
## <chr> <chr> <chr> <chr>   
## 1 Afghanistan 19 99 745/19987071   
## 2 Afghanistan 20 00 2666/20595360   
## 3 Brazil 19 99 37737/172006362   
## 4 Brazil 20 00 80488/174504898   
## 5 China 19 99 212258/1272915272  
## 6 China 20 00 213766/1280428583

table5 %>%   
 unite(new, century, year)

## # A tibble: 6 x 3  
## country new rate   
## <chr> <chr> <chr>   
## 1 Afghanistan 19\_99 745/19987071   
## 2 Afghanistan 20\_00 2666/20595360   
## 3 Brazil 19\_99 37737/172006362   
## 4 Brazil 20\_00 80488/174504898   
## 5 China 19\_99 212258/1272915272  
## 6 China 20\_00 213766/1280428583

table5 %>%   
 unite(new, century, year, sep = "")

## # A tibble: 6 x 3  
## country new rate   
## <chr> <chr> <chr>   
## 1 Afghanistan 1999 745/19987071   
## 2 Afghanistan 2000 2666/20595360   
## 3 Brazil 1999 37737/172006362   
## 4 Brazil 2000 80488/174504898   
## 5 China 1999 212258/1272915272  
## 6 China 2000 213766/1280428583

tibble(x = c("a,b,c", "d,e,f,g", "h,i,j")) %>%   
 separate(x, c("one", "two", "three"))

## Warning: Expected 3 pieces. Additional pieces discarded in 1 rows [2].

## # A tibble: 3 x 3  
## one two three  
## <chr> <chr> <chr>  
## 1 a b c   
## 2 d e f   
## 3 h i j

tibble(x = c("a,b,c", "d,e", "f,g,i")) %>%   
 separate(x, c("one", "two", "three"))

## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 1 rows [2].

## # A tibble: 3 x 3  
## one two three  
## <chr> <chr> <chr>  
## 1 a b c   
## 2 d e <NA>   
## 3 f g i

stocks <- tibble(  
 year = c(2015, 2015, 2015, 2015, 2016, 2016, 2016),  
 qtr = c( 1, 2, 3, 4, 2, 3, 4),  
 return = c(1.88, 0.59, 0.35, NA, 0.92, 0.17, 2.66)  
)

stocks %>%   
 spread(year, return)

## # A tibble: 4 x 3  
## qtr `2015` `2016`  
## <dbl> <dbl> <dbl>  
## 1 1 1.88 NA   
## 2 2 0.59 0.92  
## 3 3 0.35 0.17  
## 4 4 NA 2.66

stocks %>%   
 spread(year, return) %>%   
 gather(year, return, `2015`:`2016`, na.rm = TRUE)

## # A tibble: 6 x 3  
## qtr year return  
## <dbl> <chr> <dbl>  
## 1 1 2015 1.88  
## 2 2 2015 0.59  
## 3 3 2015 0.35  
## 4 2 2016 0.92  
## 5 3 2016 0.17  
## 6 4 2016 2.66

stocks %>%   
 complete(year, qtr)

## # A tibble: 8 x 3  
## year qtr return  
## <dbl> <dbl> <dbl>  
## 1 2015 1 1.88  
## 2 2015 2 0.59  
## 3 2015 3 0.35  
## 4 2015 4 NA   
## 5 2016 1 NA   
## 6 2016 2 0.92  
## 7 2016 3 0.17  
## 8 2016 4 2.66

treatment <- tribble(  
 ~ person, ~ treatment, ~response,  
 "Derrick Whitmore", 1, 7,  
 NA, 2, 10,  
 NA, 3, 9,  
 "Katherine Burke", 1, 4  
)

treatment %>%   
 fill(person)

## # A tibble: 4 x 3  
## person treatment response  
## <chr> <dbl> <dbl>  
## 1 Derrick Whitmore 1 7  
## 2 Derrick Whitmore 2 10  
## 3 Derrick Whitmore 3 9  
## 4 Katherine Burke 1 4

who

## # 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 Afghan~ AF AFG 1980 NA NA NA  
## 2 Afghan~ AF AFG 1981 NA NA NA  
## 3 Afghan~ AF AFG 1982 NA NA NA  
## 4 Afghan~ AF AFG 1983 NA NA NA  
## 5 Afghan~ AF AFG 1984 NA NA NA  
## 6 Afghan~ AF AFG 1985 NA NA NA  
## 7 Afghan~ AF AFG 1986 NA NA NA  
## 8 Afghan~ AF AFG 1987 NA NA NA  
## 9 Afghan~ AF AFG 1988 NA NA NA  
## 10 Afghan~ 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>

who1 <- who %>%   
 gather(new\_sp\_m014:newrel\_f65, key = "key", value = "cases", na.rm = TRUE)  
who1

## # 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

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

who2 <- who1 %>%   
 mutate(key = stringr::str\_replace(key, "newrel", "new\_rel"))  
who2

## # 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

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)

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

who4 <- who3 %>%   
 select(-new, -iso2, -iso3)

who5 <- who4 %>%   
 separate(sexage, c("sex", "age"), sep = 1)  
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

who %>%  
 gather(key, value, new\_sp\_m014:newrel\_f65, na.rm = 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)

## # 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