

Code-Along-And-Challenge-9

Ariel Quek

2023-10-16

```
knitr::opts_chunk$set(echo = TRUE)
```

Code Along 9

Part ONE: About tidy vs non-tidy data

#Tidy data

```
install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Ariel/AppData/Local/R/win-library/4.2'  
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked  
##  
## The downloaded binary packages are in  
## C:\Users\Ariel\AppData\Local\Temp\RtmpshfT7j\downloaded_packages
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
## Warning: package 'tibble' was built under R version 4.2.1
```

```
## Warning: package 'tidyr' was built under R version 4.2.2
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
## Warning: package 'purrr' was built under R version 4.2.2
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
## Warning: package 'stringr' was built under R version 4.2.2
```

```
## Warning: package 'forcats' was built under R version 4.2.3
```

```
## Warning: package 'lubridate' was built under R version 4.2.3
```

```
## — Attaching core tidyverse packages — tidyverse 2.0.0 —
## ✓ dplyr      1.1.0      ✓ readr      2.1.4
## ✓ forcats    1.0.0      ✓ stringr    1.5.0
## ✓ ggplot2    3.4.3      ✓ tibble     3.1.8
## ✓ lubridate  1.9.2      ✓ tidyr      1.3.0
## ✓ purrr      1.0.1
```

```
## — Conflicts — tidyverse_conflicts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()     masks stats::lag()
## i Use the [8];http://conflicted.r-lib.org/[8]; to force all conflict
s to become errors
```

```
tidydata <- tribble( ~country, ~year, ~cases, ~population, "Afghanistan", 1999, 745, 1998707
1,
"Afghanistan", 2000, 2666, 20595360,
"Brazil", 1999, 37737, 172006362,
"Brazil", 2000, 80488, 174504898,
"China", 1999, 212258, 1272915272,
"China", 2000, 213766, 1280428583)

tidydata
```

```
## # A tibble: 6 × 4
##   country      year  cases population
##   <chr>      <dbl> <dbl>      <dbl>
## 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
```

#Non-tidy data

```
## # A tibble: 6 × 3
##   country      year rate
##   <chr>      <dbl> <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
```

#Example of benefit of tidy data (slide 9)

```
tidydata %>%
  group_by(year) %>%
  summarize(total_cases = sum(cases))
```

```
## # A tibble: 2 × 2
##   year total_cases
##   <dbl>      <dbl>
## 1  1999      250740
## 2  2000      296920
```

#How to tidy data - Example 1 (slides 11-13)

```
tidieddata <- nontidydata %>%
  separate(rate, into = c("cases",
    "population"),
  sep = "/")
```

```
tidieddata
```

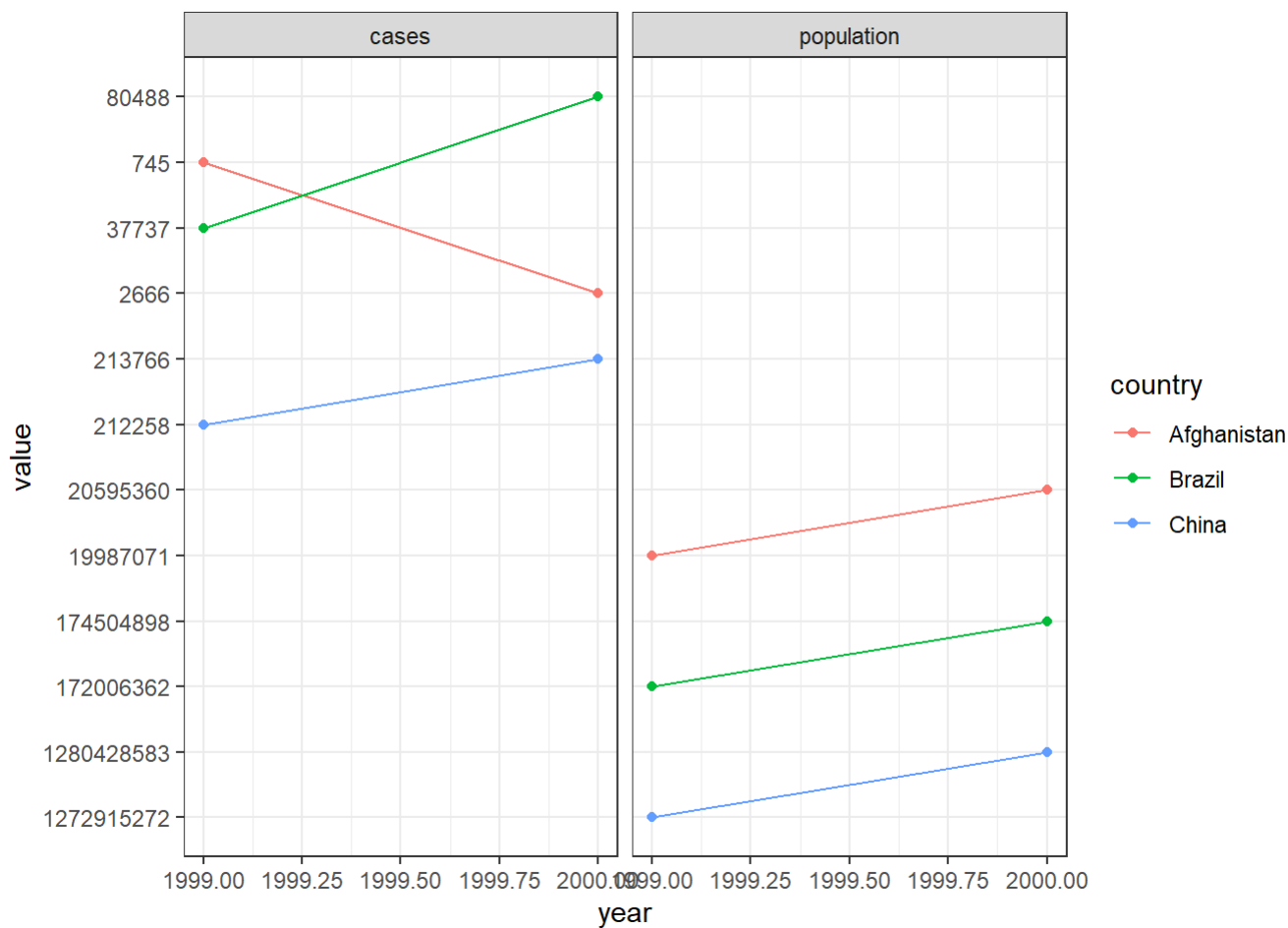
```
## # A tibble: 6 × 4
##   country      year cases population
##   <chr>      <dbl> <chr>   <chr>
## 1 Afghanistan 1999  745    19987071
## 2 Afghanistan 2000 2666    20595360
## 3 Brazil      1999 37737   172006362
## 4 Brazil      2000 80488   174504898
## 5 China       1999 212258  1272915272
##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

```
newtidieddata <- tidieddata %>%
  pivot_longer(
    cols = cases:population,
    names_to = "measurement"
  ,
  values_to = "value"
  )
```

```
newtidieddata
```

```
## # A tibble: 12 × 4
##   country      year measurement value
##   <chr>      <dbl> <chr>      <chr>
## 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
```

```
ggplot(newtidieddata) +
  aes(x=year,y=value, colour=country) +
  geom_point() +
  geom_line(aes(group = country))+
  facet_wrap(~measurement) +
  theme_bw()
```



#How to reshape data using “pivot_longer” - Example 2 (slides 14-17)

```
df <- tribble(
  ~id, ~bp1, ~bp2,
  "A", 100, 120,
  "B", 140, 115,
  "C", 120, 125)
```

```
df
```

```
## # A tibble: 3 × 3
##   id      bp1  bp2
##   <chr> <dbl> <dbl>
## 1 A      100   120
## 2 B      140   115
## 3 C      120   125
```

```
df %>%
  pivot_longer(
    cols = bp1:bp2,
    names_to = "measurement",
    values_to = "value"
  )
```

```
## # A tibble: 6 × 3
##   id    measurement value
##   <chr> <chr>      <dbl>
## 1 A      bp1         100
## 2 A      bp2         120
## 3 B      bp1         140
## 4 B      bp2         115
## 5 C      bp1         120
## 6 C      bp2         125
```

#How to reshape data using “pivot_wider” - Example 3 (slide 18)

```
newtidieddata
```

```
## # A tibble: 12 × 4
##   country    year measurement value
##   <chr>    <dbl> <chr>      <chr>
## 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
```

```
newtidieddata %>% pivot_wider(names_from="measurement",
                              values_from ="value")
```

```
## # A tibble: 6 × 4
##   country    year cases population
##   <chr>    <dbl> <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
```

#Reshaping data using “pivot_wider” - Example 4 (slide 19)

```
df <- tribble(
  ~id, ~measurement, ~value,
  "A", "bp1", 100,
  "B", "bp1", 140,
  "B", "bp2", 115,
  "A", "bp2", 120,
  "A", "bp3", 105)
df
```

```
## # A tibble: 5 × 3
##   id      measurement value
##   <chr> <chr>         <dbl>
## 1 A      bp1           100
## 2 B      bp1           140
## 3 B      bp2           115
## 4 A      bp2           120
## 5 A      bp3           105
```

```
df %>%
  pivot_wider(
    names_from = measurement,
    values_from = value)
```

```
## # A tibble: 2 × 4
##   id      bp1    bp2    bp3
##   <chr> <dbl> <dbl> <dbl>
## 1 A      100    120    105
## 2 B      140    115     NA
```

Challenge 9

Question 1

```
library(tidyverse)
billboard
```

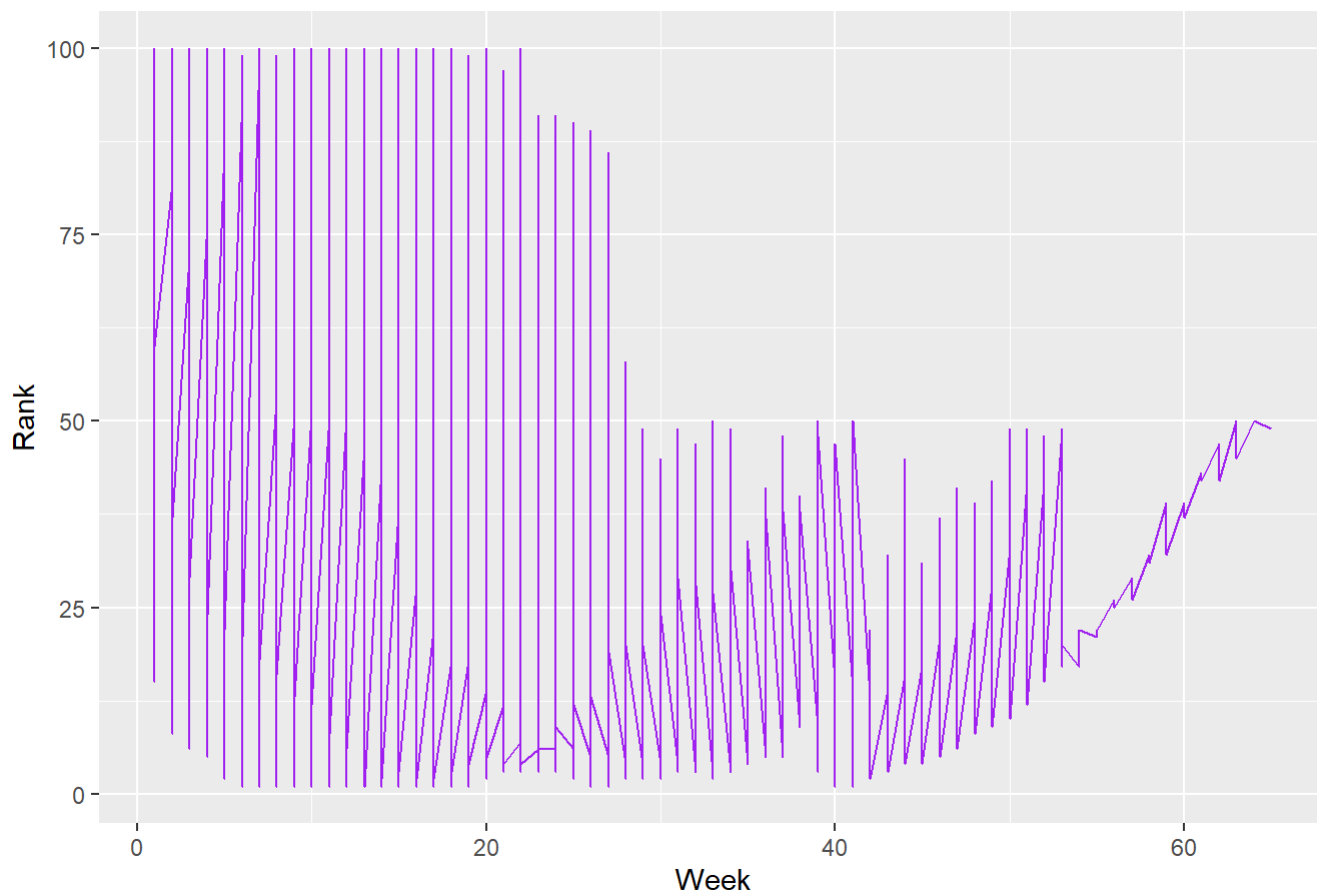
```
## # A tibble: 317 × 79
##   artist      track date.entered  wk1  wk2  wk3  wk4  wk5  wk6  wk7  wk8
##   <chr>      <chr> <date>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2 Pac      Baby... 2000-02-26    87   82   72   77   87   94   99   NA
## 2 2Ge+her    The ... 2000-09-02    91   87   92   NA   NA   NA   NA   NA
## 3 3 Doors D... Kryp... 2000-04-08    81   70   68   67   66   57   54   53
## 4 3 Doors D... Loser 2000-10-21    76   76   72   69   67   65   55   59
## 5 504 Boyz    Wobb... 2000-04-15    57   34   25   17   17   31   36   49
## 6 98^0        Give... 2000-08-19    51   39   34   26   26   19    2    2
## 7 A*Teens    Danc... 2000-07-08    97   97   96   95  100   NA   NA   NA
## 8 Aaliyah     I Do... 2000-01-29    84   62   51   41   38   35   35   38
## 9 Aaliyah     Try ... 2000-03-18    59   53   38   28   21   18   16   14
## 10 Adams, Yo... Open... 2000-08-26    76   76   74   69   68   67   61   58
## # i 307 more rows
## # i 68 more variables: wk9 <dbl>, wk10 <dbl>, wk11 <dbl>, wk12 <dbl>,
## #   wk13 <dbl>, wk14 <dbl>, wk15 <dbl>, wk16 <dbl>, wk17 <dbl>, wk18 <dbl>,
## #   wk19 <dbl>, wk20 <dbl>, wk21 <dbl>, wk22 <dbl>, wk23 <dbl>, wk24 <dbl>,
## #   wk25 <dbl>, wk26 <dbl>, wk27 <dbl>, wk28 <dbl>, wk29 <dbl>, wk30 <dbl>,
## #   wk31 <dbl>, wk32 <dbl>, wk33 <dbl>, wk34 <dbl>, wk35 <dbl>, wk36 <dbl>,
## #   wk37 <dbl>, wk38 <dbl>, wk39 <dbl>, wk40 <dbl>, wk41 <dbl>, wk42 <dbl>, ...
```

```
billboard_new <- billboard %>%
  pivot_longer(
    cols=starts_with("wk"),
    names_to="week",
    values_to="value",
    values_drop_na = TRUE) %>% mutate(week = parse_number(week))
billboard_new
```

```
## # A tibble: 5,307 × 5
##   artist track      date.entered  week value
##   <chr> <chr>      <date>      <dbl> <dbl>
## 1 2 Pac  Baby Don't Cry (Keep... 2000-02-26    1    87
## 2 2 Pac  Baby Don't Cry (Keep... 2000-02-26    2    82
## 3 2 Pac  Baby Don't Cry (Keep... 2000-02-26    3    72
## 4 2 Pac  Baby Don't Cry (Keep... 2000-02-26    4    77
## 5 2 Pac  Baby Don't Cry (Keep... 2000-02-26    5    87
## 6 2 Pac  Baby Don't Cry (Keep... 2000-02-26    6    94
## 7 2 Pac  Baby Don't Cry (Keep... 2000-02-26    7    99
## 8 2Ge+her The Hardest Part Of ... 2000-09-02    1    91
## 9 2Ge+her The Hardest Part Of ... 2000-09-02    2    87
## 10 2Ge+her The Hardest Part Of ... 2000-09-02    3    92
## # i 5,297 more rows
```

```
library(ggplot2)
# Creating a plot of rankings of tracks across time
ggplot(data=billboard_new,
  mapping = aes(x = week,
    y = value)) +
  geom_line(color="purple") +
  labs(title = "Rank of tracks across time",
    x = "Week",
    y = "Rank")
```

Rank of tracks across time



Question 2

```
cms_patient_experience
```

```
## # A tibble: 500 × 5
##   org_pac_id org_nm          measure_cd measure_title prf_rate
##   <chr>      <chr>          <chr>      <chr>          <dbl>
## 1 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP... CAHPS for MI...      63
## 2 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP... CAHPS for MI...      87
## 3 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP... CAHPS for MI...      86
## 4 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP... CAHPS for MI...      57
## 5 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP... CAHPS for MI...      85
## 6 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP... CAHPS for MI...      24
## 7 0446162697 ASSOCIATION OF UNIVERSITY PHYSI... CAHPS_GRP... CAHPS for MI...      59
## 8 0446162697 ASSOCIATION OF UNIVERSITY PHYSI... CAHPS_GRP... CAHPS for MI...      85
## 9 0446162697 ASSOCIATION OF UNIVERSITY PHYSI... CAHPS_GRP... CAHPS for MI...      83
## 10 0446162697 ASSOCIATION OF UNIVERSITY PHYSI... CAHPS_GRP... CAHPS for MI...      63
## # i 490 more rows
```



```
## A data set that collects data about patient experiences
patients <- cms_patient_experience %>%
  pivot_wider(names_from = measure_cd,
              values_from= prf_rate,
              id_cols=starts_with("org"))

##to get unique IDs
uniqueIDs <- patients %>%
  select(org_pac_id,org_nm) %>%
  distinct()
uniqueIDs
```

```
## # A tibble: 95 × 2
##   org_pac_id org_nm
##   <chr>      <chr>
## 1 0446157747 USC CARE MEDICAL GROUP INC
## 2 0446162697 ASSOCIATION OF UNIVERSITY PHYSICIANS
## 3 0547164295 BEAVER MEDICAL GROUP PC
## 4 0749333730 CAPE PHYSICIANS ASSOCIATES PA
## 5 0840104360 ALLIANCE PHYSICIANS INC
## 6 0840109864 REX HOSPITAL INC
## 7 0840513552 SCL HEALTH MEDICAL GROUP DENVER LLC
## 8 0941545784 GRITMAN MEDICAL CENTER INC
## 9 1052612785 COMMUNITY MEDICAL GROUP LLC
## 10 1254237779 OUR LADY OF LOURDES MEMORIAL HOSPITAL, INC
## # i 85 more rows
```

Thank You!