

# HW#4

Tianying Zhang

2/26/2018

## 12.6.1

```
library("tidyverse")
```

### Case Study

```
who1 <- who %>%  
  gather(new_sp_m014:newrel_f65, key = "key", value = "cases", na.rm = TRUE)  
glimpse(who1)
```

```
## Observations: 76,046  
## Variables: 6  
## $ country <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanis...  
## $ iso2 <chr> "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF", ...  
## $ iso3 <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG"...  
## $ year <int> 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, ...  
## $ key <chr> "new_sp_m014", "new_sp_m014", "new_sp_m014", "new_sp_m...  
## $ cases <int> 0, 30, 8, 52, 129, 90, 127, 139, 151, 193, 186, 187, 2...
```

```
who2 <- who1 %>%  
  mutate(key = stringr::str_replace(key, "newrel", "new_rel"))
```

```
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
```

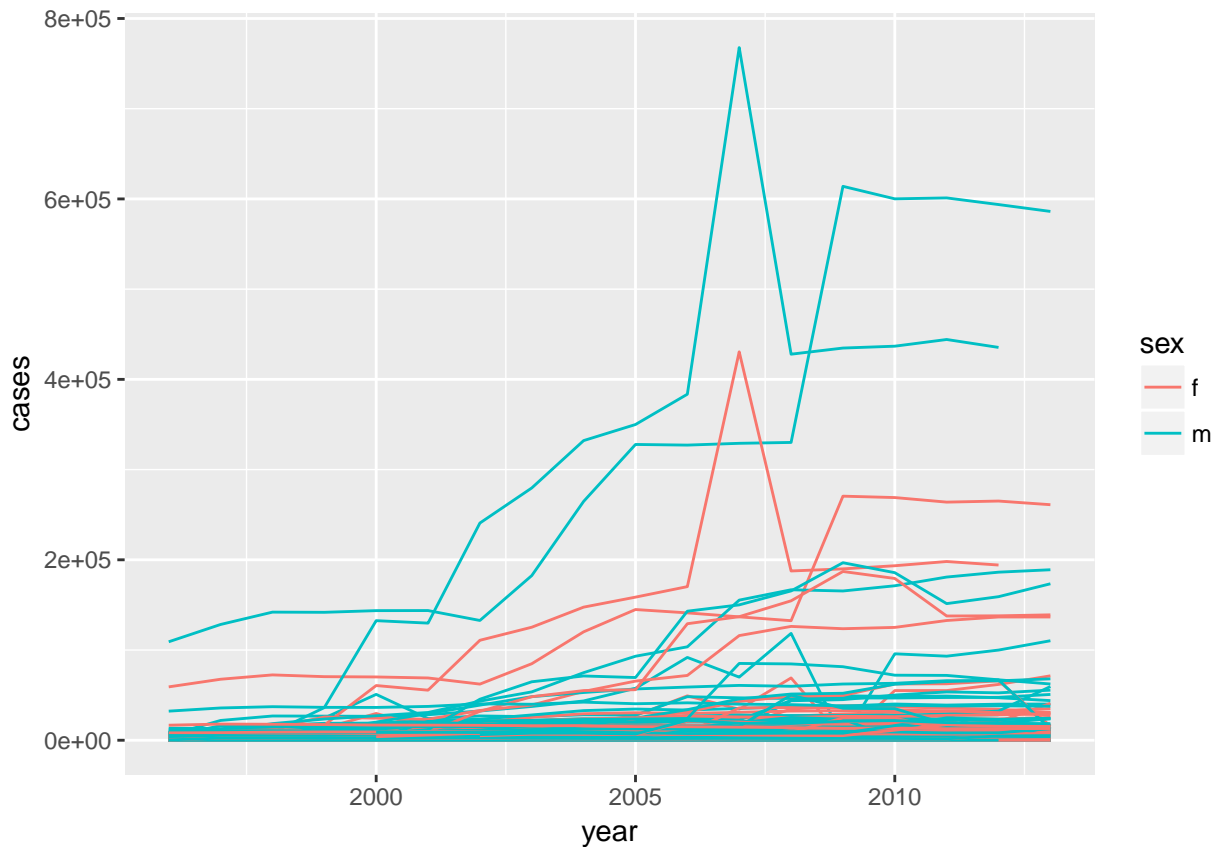
3. I claimed that iso2 and iso3 were redundant with country. Confirm this claim.

```
select(who3, country, iso2, iso3) %>%
  distinct() %>%
  group_by(country) %>%
  filter(n() > 1)
```

```
## # A tibble: 0 x 3
## # Groups:   country [0]
## # ... with 3 variables: country <chr>, iso2 <chr>, iso3 <chr>
```

4. For each country, year, and sex compute the total number of cases of TB. Make an informative visualization of the data.

```
who5 %>%
  group_by(country, year, sex) %>%
  filter(year > 1995) %>%
  summarise(cases = sum(cases)) %>%
  unite(country_sex, country, sex, remove = FALSE) %>%
  ggplot(aes(x = year, y = cases, group = country_sex, colour = sex)) +
  geom_line()
```



A small multiples plot faceting by country is difficult given the number of countries. Focusing on those countries with the largest changes or absolute magnitudes after providing the context above is another option.

## 10.5

### Tibbles

```
library("tidyverse")
```

### Exercises

5. What does `tibble::enframe()` do? When might you use it?

It converts named vectors to a data frame with names and values

```
?tibble::enframe
```

```
enframe(c(a = 1, b = 2, c = 3))
```

```
## # A tibble: 3 x 2
##   name  value
##   <chr> <dbl>
## 1 a      1.00
## 2 b      2.00
## 3 c      3.00
```

table 4 -> table 6

```
library(foreign)
library(stringr)
library(plyr)

## -----
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## -----
##
## Attaching package: 'plyr'
##
## The following objects are masked from 'package:dplyr':
##
##   arrange, count, desc, failwith, id, mutate, rename, summarise,
##   summarize
##
## The following object is masked from 'package:purrr':
##
##   compact
library(reshape2)

##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##   smiths
source("xtable.r")

# Data from http://pewforum.org/Datasets/Dataset-Download.aspx
# Load data -----

pew <- read.spss("pew.sav")

## re-encoding from CP1252
## Warning in read.spss("pew.sav"): Undeclared level(s) 2, 3, 4, 9 added in
## variable: density3
##
## Warning in read.spss("pew.sav"): Duplicated levels in factor denom:
## Electronic ministries
##
## Warning in read.spss("pew.sav"): Undeclared level(s) 1, 2, 3, 4, 5, 6, 7,
## 8, 9, 10, 11, 12, 14, 16, 23, 33 added in variable: children
##
## Warning in read.spss("pew.sav"): Undeclared level(s) 18, 19, 20, 21, 22,
## 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41,
## 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60,
## 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79,
## 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96 added in
## variable: age
```

```

pew <- as.data.frame(pew)

religion <- pew[c("q16", "reltrad", "income")]
religion$reltrad <- as.character(religion$reltrad)
religion$reltrad <- str_replace(religion$reltrad, " Churches", "")
religion$reltrad <- str_replace(religion$reltrad, " Protestant", " Prot")
religion$reltrad[religion$q16 == " Atheist (do not believe in God) "] <- "Atheist"
religion$reltrad[religion$q16 == " Agnostic (not sure if there is a God) "] <- "Agnostic"
religion$reltrad <- str_trim(religion$reltrad)
religion$reltrad <- str_replace_all(religion$reltrad, " \\(.*?\\)", "")

religion$income <- c("Less than $10,000" = "<$10k",
                    "10 to under $20,000" = "$10-20k",
                    "20 to under $30,000" = "$20-30k",
                    "30 to under $40,000" = "$30-40k",
                    "40 to under $50,000" = "$40-50k",
                    "50 to under $75,000" = "$50-75k",
                    "75 to under $100,000" = "$75-100k",
                    "100 to under $150,000" = "$100-150k",
                    "$150,000 or more" = ">150k",
                    "Don't know/Refused (VOL)" = "Don't know/refused")[religion$income]

religion$income <- factor(religion$income, levels = c("<$10k", "$10-20k", "$20-30k", "$30-40k", "$40-50k",
                                                    "$50-75k", "$75-100k", "$100-150k", ">150k", "Don't know/refused"))

counts <- count(religion, c("reltrad", "income"))
names(counts)[1] <- "religion"

raw <- dcast(counts, religion ~ income)

## Using freq as value column: use value.var to override.
raw

```

	religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k
## 1	Agnostic	27	34	60	81	76	137
## 2	Atheist	12	27	37	52	35	70
## 3	Buddhist	27	21	30	34	33	58
## 4	Catholic	418	617	732	670	638	1116
## 5	Don't know/refused	15	14	15	11	10	35
## 6	Evangelical Prot	575	869	1064	982	881	1486
## 7	Hindu	1	9	7	9	11	34
## 8	Historically Black Prot	228	244	236	238	197	223
## 9	Jehovah's Witness	20	27	24	24	21	30
## 10	Jewish	19	19	25	25	30	95
## 11	Mainline Prot	289	495	619	655	651	1107
## 12	Mormon	29	40	48	51	56	112
## 13	Muslim	6	7	9	10	9	23
## 14	Orthodox	13	17	23	32	32	47
## 15	Other Christian	9	7	11	13	13	14
## 16	Other Faiths	20	33	40	46	49	63
## 17	Other World Religions	5	2	3	4	2	7
## 18	Unaffiliated	217	299	374	365	341	528
##	\$75-100k \$100-150k >150k Don't know/refused						

```
## 1      122      109      84      96
## 2       73       59      74      76
## 3       62       39      53      54
## 4     949     792    633    1489
## 5       21       17      18     116
## 6     949     723    414    1529
## 7       47       48      54      37
## 8     131      81      78     339
## 9       15       11       6      37
## 10      69      87     151     162
## 11     939     753    634    1328
## 12      85      49      42      69
## 13      16       8       6      22
## 14      38      42      46      73
## 15      18      14      12      18
## 16      46      40      41      71
## 17       3       4       4       8
## 18     407     321     258     597
```

```
raw2 <- filter(raw, religion == "Agnostic")
```

```
raw2 %>% gather(key = "income", value = "freq", -religion)
```

```
##      religion      income freq
## 1 Agnostic      <$10k    27
## 2 Agnostic     $10-20k   34
## 3 Agnostic     $20-30k   60
## 4 Agnostic     $30-40k   81
## 5 Agnostic     $40-50k   76
## 6 Agnostic     $50-75k  137
## 7 Agnostic     $75-100k 122
## 8 Agnostic     $100-150k 109
## 9 Agnostic      >150k   84
## 10 Agnostic Don't know/refused 96
```

table 7 -> table 8

```
options(stringsAsFactors = FALSE)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:plyr':
##
##      here
## The following object is masked from 'package:base':
##
##      date
```

```
library(reshape2)
library(stringr)
library(plyr)
source("xtable.r")
```

```
raw <- read.csv("billboard.csv")
raw <- raw[, c("year", "artist.inverted", "track", "time", "date.entered",
  "x1st.week", "x2nd.week", "x3rd.week", "x4th.week", "x5th.week",
  "x6th.week", "x7th.week", "x8th.week", "x9th.week", "x10th.week",
  "x11th.week", "x12th.week", "x13th.week", "x14th.week", "x15th.week",
  "x16th.week", "x17th.week", "x18th.week", "x19th.week", "x20th.week",
  "x21st.week", "x22nd.week", "x23rd.week", "x24th.week", "x25th.week", "x26th.week", "x27th.week")]
names(raw)[2] <- "artist"

raw$artist <- iconv(raw$artist, "MAC", "ASCII//translit")
raw$track <- str_replace(raw$track, " \\(.*?\\)", "")
names(raw)[-1:5] <- str_c("wk", 1:76)
raw <- arrange(raw, year, artist, track)

long_name <- nchar(raw$track) > 20
raw$track[long_name] <- paste0(substr(raw$track[long_name], 0, 20), "...")
head(raw, 10)
```

## 2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 3	3	3	3	4	5	5	9	9	15	14	13	14	16	17
## 4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 9	43	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	wk46	wk47	wk48	wk49	wk50	wk51	wk52	wk53	wk54	wk55	wk56	wk57	wk58	wk59
## 1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 3	21	22	24	28	33	42	42	49	NA	NA	NA	NA	NA	NA
## 4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	wk60	wk61	wk62	wk63	wk64	wk65	wk66	wk67	wk68	wk69	wk70	wk71	wk72	wk73
## 1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
## 10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
##	wk74	wk75	wk76											
## 1	NA	NA	NA											
## 2	NA	NA	NA											
## 3	NA	NA	NA											
## 4	NA	NA	NA											
## 5	NA	NA	NA											
## 6	NA	NA	NA											
## 7	NA	NA	NA											
## 8	NA	NA	NA											
## 9	NA	NA	NA											
## 10	NA	NA	NA											

```

clean <- melt(raw, id = 1:5, na.rm = T)
clean$week <- as.integer(str_replace_all(clean$variable, "[^0-9]+", ""))
clean$variable <- NULL

clean$date.entered <- ymd(clean$date.entered)
clean$date <- clean$date.entered + weeks(clean$week - 1)
clean$date.entered <- NULL
clean <- rename(clean, c("value" = "rank"))
clean <- arrange(clean, year, artist, track, time, week)
clean <- clean[c("year", "artist", "time", "track", "date", "week", "rank")]

```



```
clean_out <- mutate(clean,
                     date = as.character(date))
head(clean_out,10)
```

##	year	artist	time	track	date	week	rank
## 1	2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
## 2	2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
## 3	2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
## 4	2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
## 5	2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
## 6	2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
## 7	2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
## 8	2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
## 9	2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
## 10	2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92