Assignment 4

Carlos Echeverri

2/28/2018

Section 10.5

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

By typing *?enframe*, we can see the help section that tells us that using enframe "converts named atomic vectors or lists to two-column data frames". This may be useful if we already have several vectors containing an observation with their associated value and we want to bind them together in a tibble.

Section 12.6.1

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

By typing *?tidyr::who*, we can see the description of the variables used in the data set, note that iso2 and iso3 are described as being "2 & 3 letter ISO country codes". For this reason, we can safely drop them since they are redundant with the variable that we are already using with the full country name.

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

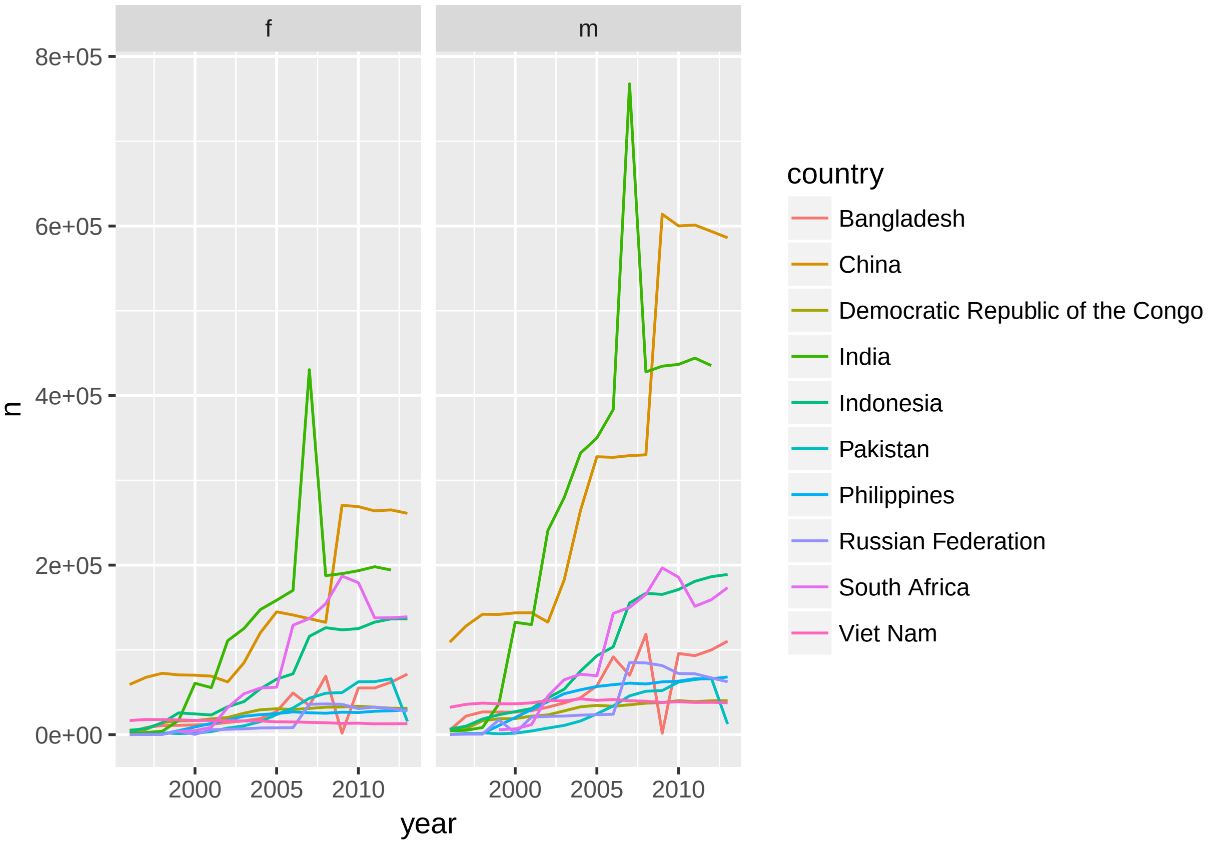
```r
library(foreign)
library(stringr)
library(plyr)
library(reshape2)
suppressMessages(library("tidyverse"))
```

```r
# Use the code provided in R for Data Science to tidy the who data set

who_tidy <- who %>%
  gather(code, value, new_sp_m014:newrel_f65, na.rm = TRUE) %>%
  mutate(code = stringr::str_replace(code, "newrel", "new_rel")) %>%
  separate(code, c("new", "var", "sexage")) %>%
  select(-new, -iso2, -iso3) %>%
  separate(sexage, c("sex", "age"), sep = 1)

# Group the data by country, create a new column that counts the number of cases
# for each of the countries. Group again by country, year and sex. Since there are many
# countries in the data set, we focus on those with the highest count of cases in order
# to be able to get more information out of the plot. We create a plot for each sex and
# give each country a different color.

who_tidy %>%
  group_by(country) %>%
  mutate(by_country = sum(value)) %>%
  group_by(country, year, sex) %>%
  filter(by_country > 900000, year > 1995 ) %>%
  count(wt = value) %>%
  ggplot(aes(year, n, color = country)) +
  geom_line() +
  facet_wrap(~ sex)
```



Using tidyverse to clean up tables

Table 4 -> Table 6

```
# Load data for table 4

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)

tab4 <- pew[c("q16", "reltrad", "income")]
tab4$reltrad <- as.character(tab4$reltrad)
```

```r
tab4$reltrad <- str_replace(tab4$reltrad, " Churches", "")
tab4$reltrad <- str_replace(tab4$reltrad, " Protestant", " Prot")
tab4$reltrad[tab4$q16 == " Atheist (do not believe in God) "] <- "Atheist"
tab4$reltrad[tab4$q16 == " Agnostic (not sure if there is a God) "] <- "Agnostic"
tab4$reltrad <- str_trim(tab4$reltrad)
tab4$reltrad <- str_replace_all(tab4$reltrad, " \\(.*?\\)", "")

tab4$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",
                 "$150,000 or more" = ">150k",
                 "Don't know/Refused (VOL)" = "Don't know/refused")[tab4$income]

tab4$income <- factor(tab4$income, levels = c("<$10k", "$10-20k", "$20-30k", "$30-40k",
                                              "$40-50k", "$50-75k",
                                              "$75-100k", "$100-150k",
                                              ">150k", "Don't know/refused"))

counts <- plyr::count(tab4, c("reltrad", "income"))
names(counts)[1] <- "religion"
tab4 <- dcast(counts, religion ~ income)
```

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

```r
tab4 <- as.tibble(tab4)

knitr::kable(head(tab4[1:7], n=10))
```

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don’t know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah’s Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

```r
#Tidy tab4 by gathering all levels of income and arranging by religion and save as tab6

tab6 <- tab4 %>% gather(key = "income", value = "freq", -religion) %>%
  arrange(religion)

knitr::kable(head(tab6, n=10))
```

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Table 7 -> Table 8

```r
# Load billboard data and create table 7

bb <- read_csv("billboard.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   artist.inverted = col_character(),
##   track = col_character(),
##   time = col_time(format = ""),
##   genre = col_character(),
##   date.entered = col_date(format = ""),
##   date.peaked = col_date(format = ""),
##   x66th.week = col_character(),
##   x67th.week = col_character(),
##   x68th.week = col_character(),
##   x69th.week = col_character(),
##   x70th.week = col_character(),
##   x71st.week = col_character(),
##   x72nd.week = col_character(),
##   x73rd.week = col_character(),
##   x74th.week = col_character(),
##   x75th.week = col_character(),
##   x76th.week = col_character()
## )
```

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

```r
tab7 <- bb %>% select(-genre, -date.peaked) %>%
  dplyr::rename(artist = artist.inverted) %>%
  arrange(artist, track) %>% mutate(track = stringr::str_trunc(track, 23, "right"))

for(i in 6:81) {
  names(tab7)[i] <- paste("wk", i-5, sep = "")
}

tab7$artist[6] <- "98^0"

knitr::kable(head(tab7[1:7], n=8))
```

year	artist	track	time	date.entered	wk1	wk2
2000	2 Pac	Baby Don’t Cry (Keep...	04:22:00	2000-02-26	87	82
2000	2Ge+her	The Hardest Part Of ...	03:15:00	2000-09-02	91	87
2000	3 Doors Down	Kryptonite	03:53:00	2000-04-08	81	70
2000	3 Doors Down	Loser	04:24:00	2000-10-21	76	76
2000	504 Boyz	Wobble Wobble	03:35:00	2000-04-15	57	34
2000	98^0	Give Me Just One Nig...	03:24:00	2000-08-19	51	39
2000	A*Teens	Dancing Queen	03:44:00	2000-07-08	97	97
2000	Aaliyah	I Don’t Wanna	04:15:00	2000-01-29	84	62

```r
# Tidy data by removing all unnecesary columns and gathering weeks in a single variable.
# Break the week column in order to keep only the number of the week. Create a formula
# that uses the date entered column and the week number in order to keep track of the
# current week.

tab8 <- bb %>% gather(key="week", value = "rank", -year, -artist.inverted,
                      -track, -time, -genre, -date.entered, -date.peaked) %>%
select(year, artist=artist.inverted, time, track, date = date.entered, week, rank ) %>%
  arrange(track) %>% filter(!is.na(rank)) %>% separate(week, into=c("A", "B", "C"),
                                                       sep=c(1, -7), convert=TRUE) %>%
  select(-A, -C) %>% dplyr::rename(week = B) %>% arrange(artist, track) %>%
  mutate(date = date + (week-1)*7 ) %>% mutate(rank = as.integer(rank)) %>%
  mutate(track = stringr::str_trunc(track, 23, "right"))

knitr::kable(head(tab8, n=15))
```

year	artist	time	track	date	week	rank
2000	2 Pac	04:22:00	Baby Don’t Cry (Keep...	2000-02-26	1	87
2000	2 Pac	04:22:00	Baby Don’t Cry (Keep...	2000-03-04	2	82
2000	2 Pac	04:22:00	Baby Don’t Cry (Keep...	2000-03-11	3	72
2000	2 Pac	04:22:00	Baby Don’t Cry (Keep...	2000-03-18	4	77
2000	2 Pac	04:22:00	Baby Don’t Cry (Keep...	2000-03-25	5	87
2000	2 Pac	04:22:00	Baby Don’t Cry (Keep...	2000-04-01	6	94
2000	2 Pac	04:22:00	Baby Don’t Cry (Keep...	2000-04-08	7	99
2000	2Ge+her	03:15:00	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	03:15:00	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	03:15:00	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	03:53:00	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	03:53:00	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	03:53:00	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	03:53:00	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	03:53:00	Kryptonite	2000-05-06	5	66