dplyr & purrr を用いたデータハンドリング

17 July, 2017

Contents

This exercise is taken from dplyr & purrr を用いたデータハンドリング. Original data is on e-Stat: 精神科病院の推計患者数，年齢階級 × 性・疾病分類（精神及び行動の障害） × 入院－外来別

1 Load Libraries

```
library(tidyverse)
```

2 Read Data

Try reading data without any option. We can see untidy data.

```
data <- readr::read_csv("j0056.csv",
                        locale = readr::locale(encoding="cp932")
                        )
knitr::kable(data[1:5, 1:3])
```

平成２６年	患者調査	平成２６年１０月
上巻第５６表	精神科病院の推計患者数，年齢階級×性・疾病分類（精神及び行動の障害）×入院－外来別	NA
NA	NA	総数
NA	NA	NA
入院	NA	NA
総数	総数	218.4

Let's skip first two lines to make it better.

```
data <- readr::read_csv("j0056.csv",
                        locale = readr::locale(encoding="cp932"),
                        skip=2)
colnames(data)
```

```
##  [1] "X1"                "X2"                "総数"
##  [4] "０歳"              "１～４"            "５～９"
##  [7] "１０～１４"        "１５～１９"        "２０～２４"
## [10] "２５～２９"        "３０～３４"        "３５～３９"
## [13] "４０～４４"        "４５～４９"        "５０～５４"
## [16] "５５～５９"        "６０～６４"        "６５～６９"
## [19] "７０～７４"        "７５～７９"        "８０～８４"
## [22] "８５～８９"        "９０歳以上"        "不詳"
## [25] "６５歳以上（再掲）" "７０歳以上（再掲）" "７５歳以上（再掲）"
```

2.1 Make tidy data

```
data %>%
    fill(X1) %>% # fill X1 with previous value, here "男"
    slice(46:63) %>%
    filter(X2 != "総数") %>%  # remove row with "総数" in column "X2"
    filter(X2 != "（再掲）精神及び行動の障害") %>%
    select(-3) %>% # delete the third column
    select(-23:-26) %>%
    rename(sex=X1, disease_name=X2) -> data
knitr::kable(data[1:5, 1:5])
```

sex	disease_name	０歳	１～４	５～９
男	血管性及び詳細不明の認知症	-	-	-
男	アルコール使用＜飲酒＞による精神及び行動の障害	-	-	-
男	その他の精神作用物質使用による精神及び行動の障害	-	-	-
男	統合失調症，統合失調症型障害及び妄想性障害	-	-	0.0
男	気分［感情］障害（躁うつ病を含む）	-	-	-

Make it tidier:

```
data %>% gather(`０歳`:`９０歳以上`, # specify columns
                key=age, # original column names will be in a new column named "age"
                value="count" # original values will be in "count"
               ) -> data
knitr::kable(data[1:10, 1:4])
```

sex	disease_name	age	count
男	血管性及び詳細不明の認知症	０歳	-
男	アルコール使用＜飲酒＞による精神及び行動の障害	０歳	-
男	その他の精神作用物質使用による精神及び行動の障害	０歳	-
男	統合失調症，統合失調症型障害及び妄想性障害	０歳	-
男	気分［感情］障害（躁うつ病を含む）	０歳	-
男	神経症性障害，ストレス関連障害及び身体表現性障害	０歳	-
男	知的障害＜精神遅滞＞	０歳	-
男	その他の精神及び行動の障害	０歳	-
女	（再掲）精神及び行動の障害	０歳	-
女	血管性及び詳細不明の認知症	０歳	-

Final modification:

```
data %>%
  rowwise() %>%
  mutate(count = ifelse(count=="-", NA, count), # create NA
         disease_name = stringr::str_trim(disease_name, side="both") # trim spaces
        ) %>%
  mutate(count = as.numeric(count)) -> data
knitr::kable(data[1:10, 1:4])
```

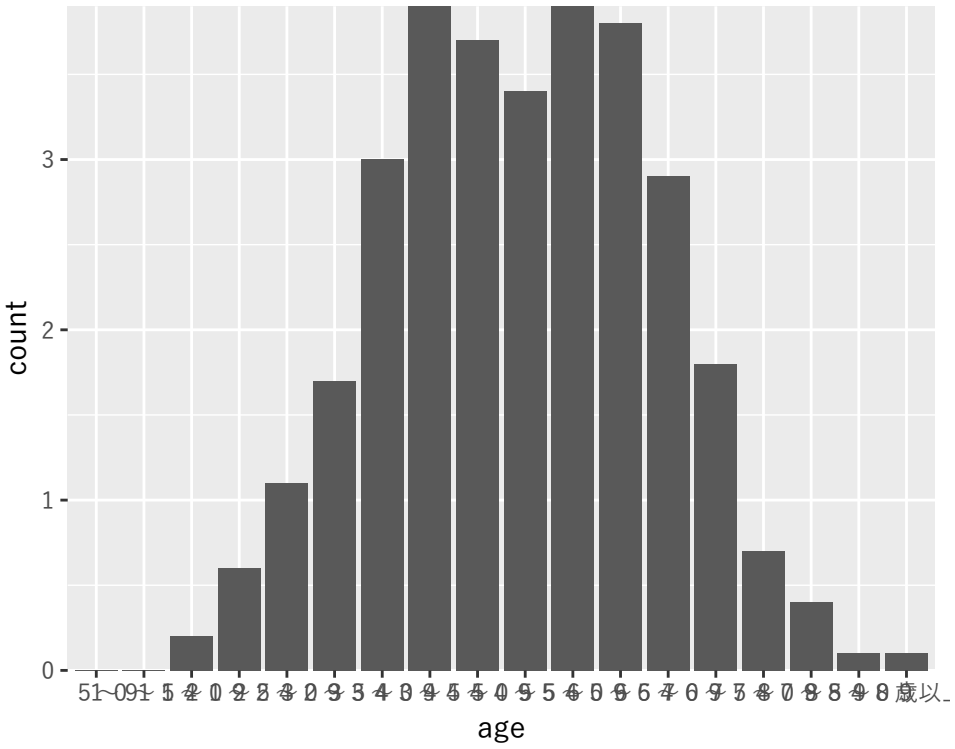
sex	disease_name	age	count
男	血管性及び詳細不明の認知症	0 歳	NA
男	アルコール使用＜飲酒＞による精神及び行動の障害	0 歳	NA
男	その他の精神作用物質使用による精神及び行動の障害	0 歳	NA
男	統合失調症，統合失調症型障害及び妄想性障害	0 歳	NA
男	気分［感情］障害（躁うつ病を含む）	0 歳	NA
男	神経症性障害，ストレス関連障害及び身体表現性障害	0 歳	NA
男	知的障害＜精神遅滞＞	0 歳	NA
男	その他の精神及び行動の障害	0 歳	NA
女	（再掲）精神及び行動の障害	0 歳	NA
女	血管性及び詳細不明の認知症	0 歳	NA

3 Transform data

```
schizo <- data %>%
           filter(disease_name == "統合失調症, 統合失調症型障害及び妄想性障害",
                  !is.na(count)) # do not take rows with NA
```

4 Visualize

```
ggplot(data=schizo) +
  geom_bar(aes(x=age, y=count), stat="identity") +
  scale_y_continuous(expand=c(0,0)) +
  theme_gray(base_family = "YuGo-Medium")
```



5 purrr

```
mapped <- data %>%
  split(list(.$disease_name, .$sex)) %>% # make list by name and sex
  map(group_by, age) %>% # applying a function to the all lists
  map(summarize, mean=mean(count, na.rm=T))
```

You can apply self-defined functions:

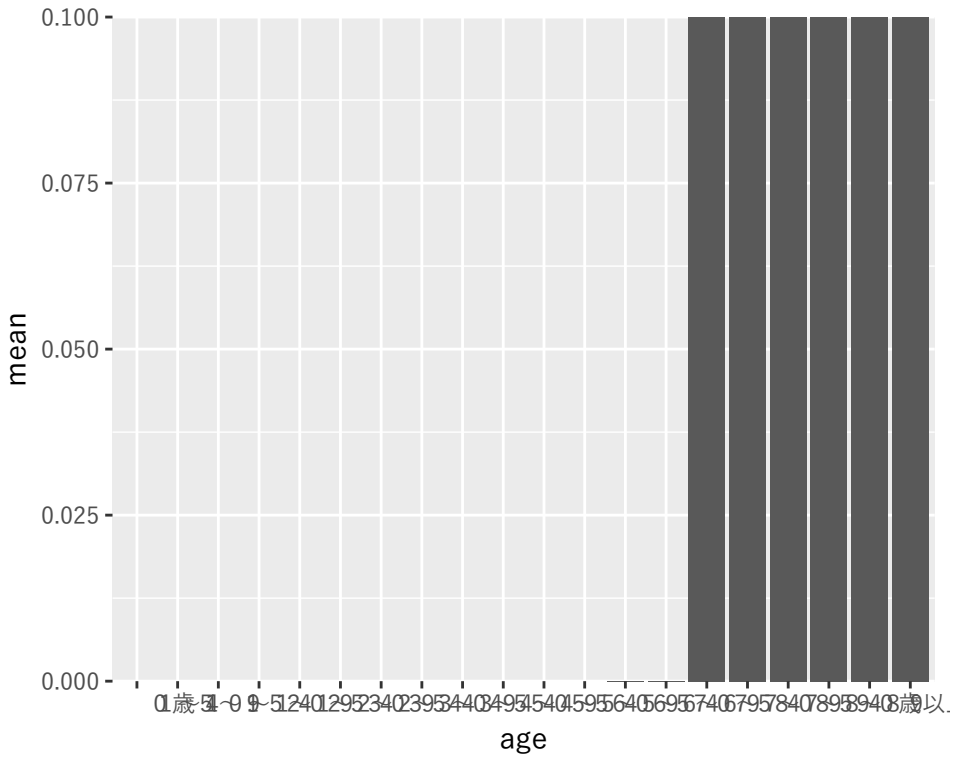
```
myhist <- function(data){
  ggplot(data) +
  geom_bar(aes(x=age, y=mean), stat="identity") +
  scale_y_continuous(expand=c(0,0)) +
  theme_gray(base_family = "YuGo-Medium")
}

mappedhist <- data %>%
  split(list(.$disease_name, .$sex)) %>% # make list by name and sex
  map(group_by, age) %>% # applying a function to the all lists
  map(summarize, mean=mean(count, na.rm=T)) %>%
  map(myhist)
names(mappedhist)
```

```
##  [1] "（再掲）精神及び行動の障害.女"
##  [2] "アルコール使用＜飲酒＞による精神及び行動の障害.女"
##  [3] "その他の精神作用物質使用による精神及び行動の障害.女"
##  [4] "その他の精神及び行動の障害.女"
##  [5] "気分［感情］障害（躁うつ病を含む）.女"
##  [6] "知的障害＜精神遅滞＞.女"
```

```
##  [7] "神経症性障害, ストレス関連障害及び身体表現性障害.女"
##  [8] "統合失調症, 統合失調症型障害及び妄想性障害.女"
##  [9] "血管性及び詳細不明の認知症.女"
## [10] "（再掲）精神及び行動の障害.男"
## [11] "アルコール使用＜飲酒＞による精神及び行動の障害.男"
## [12] "その他の精神作用物質使用による精神及び行動の障害.男"
## [13] "その他の精神及び行動の障害.男"
## [14] "気分［感情］障害（躁うつ病を含む）.男"
## [15] "知的障害＜精神遅滞＞.男"
## [16] "神経症性障害, ストレス関連障害及び身体表現性障害.男"
## [17] "統合失調症, 統合失調症型障害及び妄想性障害.男"
## [18] "血管性及び詳細不明の認知症.男"
```

You can take out any figure:

```
mappedhist$血管性及び詳細不明の認知症.男
```



You can make a list of analysis:

```
data %>% split(.$disease_name) %>%
  map(~lm(count ~ sex, data=.)) %>%
  map(summary) %>%
  map("coefficients")
```

6 Map family

- Change return values
- map_lg()
- map_chr()
- map_int()

- `map_dbl()`
- map multiple lists
- `map2()`
- `map3()`
- `map_n()`