R
R.package

dplyr

---

library(dplyr)

%>% でパイプ処理

・左の命令の結果を、右の命令に受け渡す
・入れ子型に丸かっこで何重に入れこまなくても

```
mean(head(women$height))
```

・ステップごとに分けて書けるので、見やすい。

```
head(women$height) %>% mean
```

■キーボード・ショートカット Ctrl+Shift+M

filter()

■レコードにフィルターをかけて、必要なレコードだけを選び出す

■複数の条件は、カッコ内に条件を並べる。

```
filter(doc_id == "text1" , sentence_id == 2)
```

■ %in% で左右に並べることで、左の中のレコードで右側のどれかに該当するものを選ぶ

・左側がデータフレームで、その中の見出しに含まれるもののうち、特定の該当するものについて、ベクトルにまとめてあるものを右側にして、該当するものだけのレコードを抽出する。

```
df %>% filter(id %in% hit)
```

■該当しないものを選ぶ（上の「否定」）：複数のものを除く場合

・ 事前に除くもののリストを作っておいて

```
nozoku <- c("iranai","hosikunai")
```

・それを否定で指定する

```
filter(!( 全体リスト %in% nozoku))
```

```
df %>% filter(!(id %in% hit))
```

・hit の部分は以下のようでも可

```
filter(!(ID %in% c(1805, 1816, 1817, 1866, 1916, 1925, 1927, 1936, 1938, 2022, 2045, 2049)))
```

■特定の行を削除するには != で該当しないものを残せばよい

```
NP.dat.jp2b <- NP.dat.jp %>% filter(Group != "Am")
```

・しかし、これだと、Group に Am というレベルが残ったままになる。
・使ってないレベルを削除するのは droplevels()

```
NP.dat.jp2c <- droplevels(NP.dat.jp2)
```

■ NICER1.1 の言語特徴量データフレームから、特定のトピックだけを抜き出す

select()

■特定のカラムだけ抽出

```
errdat5 <- select(errdat4, M.OTHER, R.OTHER, U.OTHER)
```

■ errdat4 の中から、M.OTHER, R.OTHER, U.OTHER の三種類のカラムだけ選ぶ

■特定のカラムを含まない残りを抽出

・含めたくないものの列名にマイナスをつけて並べる

```
errdat5 <- select(errdat4, -M.OTHER, -R.OTHER, -U.OTHER)
```

mutate()

■ mutate(付け足すカラム名 = 変換操作)

・結果を保存するには、普通に左辺に代入すればよい

```
ToothGrowthResult <- ToothGrowth %>% mutate(result = len * dose)
```

■定型文字列を一列つけるには

```
mutate(Lang = "JP")
```

・操作によるが、単につけるだけなら、paste を使ってもできる。

```
jp2gram$bigram <- paste(jp2gram$first, jp2gram$second)
```

■条件によって、文字列を変えるには if_else を使う

・mutate(Lang = if_else(判断もとの列名 == 条件 , 入力文字 , それ以外の場合に入力する文字)
・例

```
mutate(Lang = if_else(year=="jp", "L1", "L2"))
```

・year の列が jp だったら Lang に L1 といれて、jp 以外だったら L2 と入れる

・例：正規表現を使った例
・ str_detect() を使う

```
PT.dat2 %>% mutate(Year = if_else(str_detect(KID, "^19"), 3, if_else(str_detect(KID, "^20"), 2, 1)))
```

・学籍番号が 19 で始まる場合は、Year=3
・学籍番号が 20 で始まる場合は、Year=2
・その他の場合（学籍番号が 21 で始まる）は、Year=1

■付け足さずに、置き換えることもできる。（データフレーム内の置き換え）

```
SbyS.datN <- SbyS.dat %>% mutate(Year=gsub("NS", "4", Year))
```

・Year というカラムに、2 と 3 と NS と入っているところ、NS を 4 に置き換える

■行で集計する apply とオプションの 1

・例：3 列目から 62 列目までの数値を合計 (sum) して、total というカラムを最後に追加する。

```
data %>% mutate(total = apply(data[, c(3:62)], 1, sum))
```

top_n(上位何位まで , カラム名)

・順位を指定しても、タイがある場合は、該当するもの全部
・上位を出してくれるが、並べ替えはしてくれない（並べ替えは、下の arrange()）

■下位から選ぶ場合は、マイナスをつける

arrange()

■昇順

■降順は arrange(desc())

rename()

■カラム名を変更

```
rename( データ , 新しい名前 = 元の名前 )
```

・同時に複数変更

```
rename( データ , 新しい名前 = 元の名前 , 新しい名前２ = 元の名前２ )
```

・変えたいところだけ指定すればよい

・標準の colnames() だと、全部指定しないといけない

relocate()

■カラムを並べる順序を変える

・順番に前から指定する（指定されないものはもとのまま）

```
relocate( データフレーム , 一番前に来る見出し )
relocate( データフレーム , 一番前に来る見出し , 二番目 )
relocate( データフレーム , 一番前に来る見出し , 二番目 , 三番目 )
```

・特定の場所に入れる
・あるものの前

```
relocate( データフレーム , 動かす見出し , .before= あるもの )
```

・あるものの後

```
relocate( データフレーム , 動かす見出し , .after= あるもの )
```

full_join()

■二つのデータフレームを一つに統合する。

・統合するキーを設定して、それをもとに、それぞれ該当がないところは NA とする。

■二種類の頻度表を合わせて比較できるようにする

データフレームの縦横結合

■データフレームの縦結合： dplyr::bind_rows

・2 つ以上も OK

■データフレームの横結合： dplyr::bind_cols

・2 つ以上も OK

select()

・ カラムの選択

■オプションとして

■ matches()

・正規表現で列名の指定ができる。

■ starts_with()

■ ends_with()

■ contains()

pull()

・選択したカラムの値をベクトルで取り出す

group_by()

・グループごとにまとめて集計する
・二つ（以上）指定して、入れ子にすることも可能。
　・話し言葉か書き言葉かという mode で分けてから、学年 year で分けて、N から IPSyn までの、平均と SD を出す。

■ ungroup(**データ**)

・分類し終わったら、ungroup しておかないと、いつまでも分けられた状態のデータになったまま。

summarise(**見出し** = **命令** ())

・集計する

・
　・その前に、na.omit() しておく。
　・もしくは、オプション na.rm=T をつけて

```
summarise( 平均 =mean(Score, na.rm = T), 標準偏差 =sd(Score, na.rm = T))
```

　・関係ないカラムは、select(- 不要カラム名 , - 不要カラム名) で除いておく

・参考

https://izunyan.github.io/gisho12/summarise.html
https://www.jaysong.net/RBook/datahandling2.html

使用例

■月ごとに分けて概要をまとめる。

・group_by() %>% summarise()

summarise_each()