stringr

- stringi のラッパー
- tidyverse に含まれている

■正規表現のオプション

- ignore_case=T
- multiline=T

■ str_sub() 文字列の抽出

```
mutate(KID = str_sub(DateID, start= 6, end=9)
```

- "2107_1901" の 6 文字目から 9 文字目までを抜き出す。結果 1901

■ str_starts() 文字列の検索（始まる文字列）

```
sp.dat.long.score %>% dplyr::filter(str_starts(name, "Mostafa"))
```

・long フォーマットのデータのうち、見出し name のところで、
・"Mostafa" で始まる文字列からなる項目を含む行のみを選び出す。

■ str_ends() 文字列の検索（終わる文字列）

```
sp.dat.long.score %>% dplyr::filter(str_ends(name, "_DET"))
```

■ str_c() 文字列の結合

・例：ID = str_c(Lang,Year,PID,SID, sep="_"
　・Lang, Year, PID, SID をアンダースコアでつないで、新しく ID という文字列にする

■ str_which() 文字列がある行番号を調べる

```
str_which( カラム名 ," 文字列 ")
```

■ str_detect() 該当する文字列があるかどうか調べる

```
str_detect( データ , " 正規表現 ")
```

・subset() と合わせて使うと便利
　・データフレーム中の特定の列に「ある種の文字列」があるかどうかを調べて、その文字列を含む行だけを選び出す。
　　・「ある種の文字列」の例：小文字の連続で書かれている「単語」が複数あるもの

・filter と合わせて使う例
  ・ KID というカラムで、21 で始まる文字列の行だけ選ぶ
    ・ ^ が文字列の始まり指定

```
dplyr::filter(str_detect(KID, "^21"))
```

■ str_extract() **指定したパターンが該当した文字列を抽出する**

・正規表現で複数のパターンの文字列が該当する場合、個々に該当したパターンを出力

■ str_replace(**文字列** , **置換対象表現** , **置換後表現**)

■ str_remove(**文字列** , **削除表現**)

・ 上の置換で、何もなしで置換と同等

■ str_remove_all(**文字列** , **削除表現**)

・ 文字列中に出現しているすべての該当パタンを削除
  ・ _all なしだと、最初のものだけしか処理しない

■ str_count(**文字列** , '\\w+') **単語数のカウント**

・一文字以上の英字の連続（\w+）の数を数える

https://www.statology.org/word-count-in-r/

References **参考サイト**

https://heavywatal.github.io/rstats/stringr.html
https://evoldyn.gitlab.io/evomics-2018/ref-sheets/R_strings.pdf