

もうこわくない！Rで読む ひどい列名のExcelファイル



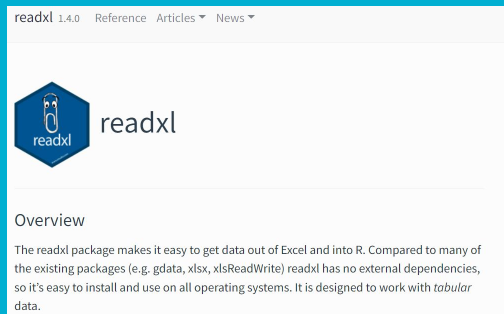
やわらかクジラ



:@matsuchiy

同人活動 (サークル名: ヤサイゼリー)

- 技術書典9にて頒布^[1]
- RでのExcelファイルの読み書き
 - readxlパッケージによる1つ～大量のxlsxファイル
 - csv × windowsの文字化けのつらみへの対処



[1] pdf: <https://techbookfest.org/product/4794168259903488?productVariantID=5913872206659584>, html: https://izunyan.github.io/excel_r/

ひどい列名の例(架空)

	A	B	C	D	E	F	G	H	I
1	<u>Species</u>	種 類	※島の名前	①クチバシ 長さ(mm)	②クチバシ 大きさ(mm)	翼:長さ(mm)	■体重 単位はg	♂(男)	2007~2009
2	<u>Adelie</u>	アデリー	<u>Torgersen</u>	39.1	18.7	181	3750	male	2007
3	<u>Adelie</u>	アデリー	<u>Torgersen</u>	39.5	17.4	186	3800	female	2007
4	<u>Adelie</u>	アデリー	<u>Torgersen</u>	40.3	18	195	3250	female	2007
5	<u>Adelie</u>	アデリー	<u>Torgersen</u>						2007
6	<u>Adelie</u>	アデリー	<u>Torgersen</u>	36.7	19.3	193	3450	female	2007
7	<u>Adelie</u>	アデリー	<u>Torgersen</u>	39.3	20.6	190	3650	male	2007
8	<u>Adelie</u>	アデリー	<u>Torgersen</u>	38.9	17.8	181	3625	female	2007
9	<u>Adelie</u>	アデリー	<u>Torgersen</u>	39.2	19.6	195	4675	male	2007

- 全角半角混在
- 不要なスペース
- 特殊記号
- 数字始まり



読み込んでみる

パッケージ名::関数など、
の書き方で直接読みだせる

```
> df <-  
+   readxl::read_xlsx("data/ペンギン (ひどい列名) ver.xlsx")  
> df  
# A tibble: 344 x 9  
  Species `種類` `※島の名前` `①クチバシ 長さ (~` `②クチバシ_大きさ (~`  
    <chr>      <chr>      <chr>          <dbl>          <dbl>  
1 Adelie     アデリー   Torgersen      39.1           18.7  
2 Adelie     アデリー   Torgersen      39.5           17.4  
3 Adelie     アデリー   Torgersen      40.3           18  
4 Adelie     アデリー   Torgersen      NA             NA  
5 Adelie     アデリー   Torgersen      36.7           19.3  
6 Adelie     アデリー   Torgersen      39.3           20.6  
7 Adelie     アデリー   Torgersen      38.9           17.8  
8 Adelie     アデリー   Torgersen      39.2           19.6  
9 Adelie     アデリー   Torgersen      34.1           18.1  
10 Adelie    アデリー   Torgersen      42            20.2  
# ... with 334 more rows, and 4 more variables: 翼:長さ(mm) <dbl>,  
#   体重 単位はg <dbl>, <U+329B><U+329A> <chr>, 2007~2009 <dbl>
```

```
> glimpse(df)  
Rows: 344  
Columns: 9  
$ Species      <chr> "Adelie", "Adelie",  
$ `種類`       <chr> "アデリー", "アデリー",  
$ `※島の名前` <chr> "Torgersen", "Torgersen",  
$ `①クチバシ 長さ (mm)` <dbl> 39.1, 39.5, 40.3, NA,  
$ `②クチバシ_大きさ (mm)` <dbl> 18.7, 17.4, 18.0, NA,  
$ `翼:長さ(mm)` <dbl> 181, 186, 195, NA,  
$ `体重 単位はg` <dbl> 3750, 3800, 3250, NA,  
$ `性別`       <chr> "male", "female",  
$ `2007~2009` <dbl> 2007, 2007, 2007, 2008
```

読めるが扱いが不便に

データフレーム %>%

適用する関数(対象の変数)

- “ ” (ダブルクォーテーション)などで囲まないと扱えない

```
> df %>% select(①クチバシ 長さ (mm) )
Error: unexpected input in "df %>% select(①"
> df %>% select("①クチバシ 長さ (mm) ")
# A tibble: 344 x 1
  `①クチバシ 長さ (mm)`
      <dbl>
1         39.1
2         39.5
3         40.3
4          NA
5         36.7
6         39.3
7         38.9
8         39.2
9         34.1
10        42
# ... with 334 more rows
```

地道にrenameすれば扱えるが大変

```
> df %>%
+   rename(species = species,
+           item1_bill_length_mm = "①クチバシ 長さ (mm) ",
+           sex = "②♀♂")
# A tibble: 344 x 9
#   species `種類` `※島の名前` item1_bill_length_mm `②クチバシ_大きさ (mm~
#   <chr>   <chr>   <chr>               <dbl>               <dbl>
1 Adelie アデリー Torgersen          39.1             18.7
2 Adelie アデリー Torgersen          39.5             17.4
3 Adelie アデリー Torgersen          40.3              18
4 Adelie アデリー Torgersen           NA              NA
5 Adelie アデリー Torgersen          36.7             19.3
6 Adelie アデリー Torgersen          39.3             20.6
7 Adelie アデリー Torgersen          38.9             17.8
8 Adelie アデリー Torgersen          39.2             19.6
9 Adelie アデリー Torgersen          34.1             18.1
10 Adelie アデリー Torgersen          42              20.2
# ... with 334 more rows, and 4 more variables: 翼:長さ(mm) <dbl>,
#   ■体重 単位はg <dbl>, sex <chr>, 2007~2009 <dbl>
```

janitor::clean_names()で

- 引数に
 case = "old_janitor"
- 特殊記号は **x**, スペースが
 _ に変換
- 数字始まりも **x** 始まりに

```
> df %>%  
+   janitor::clean_names(case = "old_janitor")  
# A tibble: 344 x 9  
  species 種類      x_島の名前 x_クチバシ_長さ_mm x_クチバシ_大きさ_m~  
  <chr>    <chr>    <chr>          <dbl>          <dbl>  
1 Adelie   アデリー Torgersen      39.1           18.7  
2 Adelie   アデリー Torgersen      39.5           17.4  
3 Adelie   アデリー Torgersen      40.3           18  
4 Adelie   アデリー Torgersen      NA             NA  
5 Adelie   アデリー Torgersen      36.7           19.3  
6 Adelie   アデリー Torgersen      39.3           20.6  
7 Adelie   アデリー Torgersen      38.9           17.8  
8 Adelie   アデリー Torgersen      39.2           19.6  
9 Adelie   アデリー Torgersen      34.1           18.1  
10 Adelie  アデリー Torgersen      42             20.2  
# ... with 334 more rows, and 4 more variables: 翼_長さ_mm <dbl>,  
#   x_体重_単位はg <dbl>, x_u_329b_u_329a <chr>, 2007_2009 <dbl>
```

(参考)疫学者のためのRハンドブックでも

The Epidemiologist R Handbook

Table of contents

About this book

1 Editorial and technical notes

2 Download handbook and data

Basics

3 R Basics

4 Transition to R

5 Suggested packages

6 R projects

7 Import and export

Data Management


8 **Cleaning data and core functions**

9 Working with dates

10 Characters and strings

8 Cleaning data and core functions

Date of Onset	Sex	Age
1/1/1965	M	24 years
15 March 1994	NA	16 months
13 Dec. 1989	Fem	29
25/6/2001	F	3



date_onset	sex	age_years
1965-01-01	Male	24.00
1994-03-15	Missing	1.33
1989-12-13	Female	29.00
2001-06-25	Female	3.00

This page demonstrates common steps used in the process of "cleaning" a dataset, and also explains the use of many essential R data management functions.

To demonstrate data cleaning, this page begins by importing a raw case list dataset, and proceeds step-by-step through the cleaning process. In the R code, this manifests as a "pipe" chain, which references the "pipe" operator `%>%` that passes a dataset from one operation to the next.

Core functions

This handbook emphasizes use of the functions from the **tidyverse** family of R packages. The essential R functions demonstrated in this page are listed below.

On this page

- 8 Cleaning data and core functions
- 8.1 Cleaning pipeline
- 8.2 Load packages
- 8.3 Import data
- 8.4 Column names
- 8.5 Select or re-order columns
- 8.6 Deduplication
- 8.7 Column creation and transformation
- 8.8 Re-code values
- 8.9 Numeric categories
- 8.10 Add rows
- 8.11 Filter rows
- 8.12 Row-wise calculations
- 8.13 Arrange and sort

Function	Utility	Package
<code>%>%</code>	"pipe" (pass) data from one function to the next	magrittr
<code>mutate()</code>	create, transform, and re-define columns	dplyr
<code>select()</code>	keep, remove, select, or re-name columns	dplyr
<code>rename()</code>	rename columns	dplyr
<code>clean_names()</code>	standardize the syntax of column names	janitor
<code>as.character()</code> , <code>as.numeric()</code> , <code>as.Date()</code> , etc.	convert the class of a column	base R
<code>across()</code>	transform multiple columns at one time	dplyr

stringi::stri_trans_nfkc()で

- 全角が半角に変換
- 特殊記号のいくつかも文字に

rename_with(~適用する関数(.x),
everything())
で全ての変数に関数を適用

```
> df %>%
+   rename_with(~stringi::stri_trans_nfkc(.x),
+               everything())
# A tibble: 344 x 9
  Species `種 類` `※島の名前` `1クチバシ 長さ(mm)` `2クチバシ_大きさ(mm)`
  <chr>    <chr>    <chr>                <dbl>                <dbl>
1 Adelie アデリー Torgersen             39.1                 18.7
2 Adelie アデリー Torgersen             39.5                 17.4
3 Adelie アデリー Torgersen             40.3                 18
4 Adelie アデリー Torgersen             NA                    NA
5 Adelie アデリー Torgersen             36.7                 19.3
6 Adelie アデリー Torgersen             39.3                 20.6
7 Adelie アデリー Torgersen             38.9                 17.8
8 Adelie アデリー Torgersen             39.2                 19.6
9 Adelie アデリー Torgersen             34.1                 18.1
10 Adelie アデリー Torgersen             42                    20.2
# ... with 334 more rows, and 4 more variables: 翼:長さ(mm) <dbl>,
#   体重 単位はg <dbl>, 女男 <chr>, 2007~2009 <dbl>
```

組み合わせると全部解決！

```
> df %>%
+   rename_with(~stringi::stri_trans_nfkc(.x),
+               everything()) %>%
+   janitor::clean_names(case = "old_janitor")
# A tibble: 344 x 9
  species 種類    x_島の名前 x1クチバシ_長さ~ x2クチバシ_大き~ 翼_長さ_mm
  <chr>    <chr>    <chr>          <dbl>          <dbl>      <dbl>
1 Adelie アデリー Torgersen      39.1           18.7       181
2 Adelie アデリー Torgersen      39.5           17.4       186
3 Adelie アデリー Torgersen      40.3           18         195
4 Adelie アデリー Torgersen      NA             NA         NA
5 Adelie アデリー Torgersen      36.7           19.3       193
6 Adelie アデリー Torgersen      39.3           20.6       190
7 Adelie アデリー Torgersen      38.9           17.8       181
8 Adelie アデリー Torgersen      39.2           19.6       195
9 Adelie アデリー Torgersen      34.1           18.1       193
10 Adelie アデリー Torgersen      42             20.2       190
# ... with 334 more rows, and 3 more variables: x_体重_単位はg <dbl>,
#   女男 <chr>, x2007_2009 <dbl>
```

開始行が1行目じゃないファイル

	A	B	C	D	E	F	G	H	I
1									
2	ここに説明とか書いてあるファイル読むのつらいのです								
3									
4	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	種類
5	Adelie	Torgersen	39.1	18.7	181	3750	male	2007	アデリー
6	Adelie	Torgersen	39.5	17.4	186	3800	female	2007	アデリー
7	Adelie	Torgersen	40.3	18	195	3250	female	2007	アデリー
8	Adelie	Torgersen						2007	アデリー
9	Adelie	Torgersen	36.7	19.3	193	3450	female	2007	アデリー
10	Adelie	Torgersen	39.3	20.6	190	3650	male	2007	アデリー
11	Adelie	Torgersen	38.9	17.8	181	3625	female	2007	アデリー
12	Adelie	Torgersen	39.2	19.6	195	4675	male	2007	アデリー
13	Adelie	Torgersen	34.1	18.1	193	3475		2007	アデリー
14	Adelie	Torgersen	42	20.2	190	4250		2007	アデリー
15	Adelie	Torgersen	37.8	17.1	186	3300		2007	アデリー
16	Adelie	Torgersen	37.8	17.3	180	3700		2007	アデリー
17	Adelie	Torgersen	41.1	17.6	182	3200	female	2007	アデリー

```
> read_xlsx("data/ペンギン (3行空き).xlsx")
New names:
* --> ...1
* --> ...3
* --> ...4
* --> ...5
* --> ...6
* ...
# A tibble: 346 x 9
  ...1   ...3   ...4   ...5   ...6   ...7   ...8   ...9
  <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
1 NA    NA    NA    NA    NA    NA    NA    NA
2 species island bill_~ bill_~ flipp~ body~ sex  year 種類
3 Adelie Torgersen 39.1 18.7 181 3750 male 2007 アデリー
4 Adelie Torgersen 39.5 17.4 186 3800 fema~ 2007 アデリー
5 Adelie Torgersen 40.3 18 195 3250 fema~ 2007 アデリー
6 Adelie Torgersen NA NA NA NA NA 2007 アデリー
7 Adelie Torgersen 36.7 19.3 193 3450 fema~ 2007 アデリー
8 Adelie Torgersen 39.3 20.6 190 3650 male 2007 アデリー
9 Adelie Torgersen 38.9 17.8 181 3625 fema~ 2007 アデリー
10 Adelie Torgersen 39.2 19.6 195 4675 male 2007 アデリー
# ... with 336 more rows
```

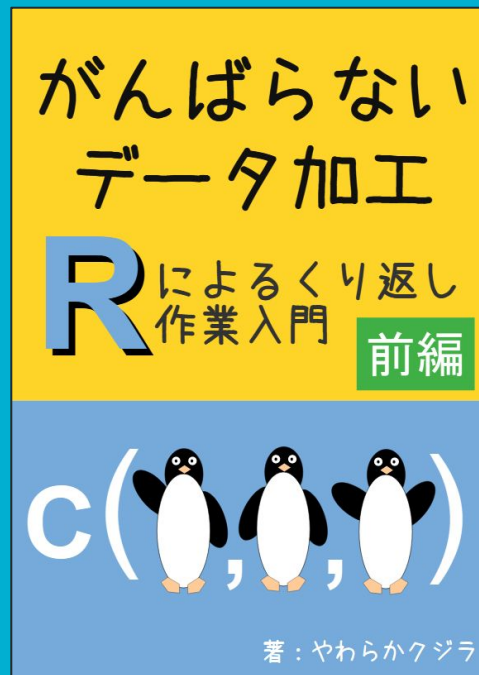
read_xlsx()のskip引数

```
> readxl::read_xlsx("data/ペンギン (3行空き) .xlsx", skip = 3)
# A tibble: 344 x 9
  species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g
  <chr>    <chr>      <dbl>         <dbl>         <dbl>         <dbl>
1 Adelie  Torger~      39.1          18.7          181          3750
2 Adelie  Torger~      39.5          17.4          186          3800
3 Adelie  Torger~      40.3           18          195          3250
4 Adelie  Torger~      NA           NA           NA           NA
5 Adelie  Torger~      36.7          19.3          193          3450
6 Adelie  Torger~      39.3          20.6          190          3650
7 Adelie  Torger~      38.9          17.8          181          3625
8 Adelie  Torger~      39.2          19.6          195          4675
9 Adelie  Torger~      34.1          18.1          193          3475
10 Adelie Torger~      42           20.2          190          4250
# ... with 334 more rows, and 3 more variables: sex <chr>, year <dbl>,
#   種類 <chr>
```

(参考) 同人活動

(サークル名: ヤサイゼリー)

- 技術書典12にて頒布^[1]
- 前編では, Rの基本知識とdplyrの基本動詞を解説
 - ヘルパー関数, `rename_with`, `across`が分かる人には不要な本
 - 逆に知らない人には効率化にすごく役立つはず
- 第98回R勉強会@東京(#TokyoR)の初心者セッションで発表^[2]



[1] pdf: <https://techbookfest.org/product/5161487259664384?productVariantID=5672571053801472>, html: <https://izunyan.github.io/gisho12/>

[2] <https://tokyor.connpass.com/event/244200/>

Rと仲良くなるヒント(自分の場合)

- 自分のデータを読み込むのは, 読み込まずにいられないぐらいに dplyrの基本動詞になれてから
- まずはサンプルデータで以下の5つの動詞
 - 列(変数, カラム)を選ぶ: **select**
 - 変数名を変更する : **rename**
 - 行(ケース)を選ぶ : **filter**
 - 新しい変数(列)の作成 : **mutate**
 - 要約値を作る : **summarise**
- RStudio開いたら `library(tidyverse)`

本発表のコードはこれが前提

まとめ

- もうひどい列名に直面してもこわくない
- Rは統計解析環境だけでなく前処理の便利なツールたくさん
- tidyverseと仲良くなって豊かな前処理ライフを

Enjoy!

