

tidyverse এর মাধ্যমে ডেটা আমদানি : : চিট শিট



readr এর মাধ্যমে ট্যাবুলার ডেটা পড়া

`read_*`(file, col_names = TRUE, col_types = NULL, col_select = NULL, id = NULL, locale, n_max = Inf, skip = 0, na = c("", "NA"), guess_max = min(1000, n_max), show_col_types = TRUE) See `?read_delim`

`read_delim`("file.txt", delim = "|") যেকোনো ডিলিমিটার সহ ফাইল পড়ুন। যদি কোন ডিলিমিটার না উল্লেখ থাকে, তাহলে এটা স্বয়ংক্রিয় অনুমান করবে।
file.txt তৈরি করার জন্য লিখুন: `write_file("A|B|C\n1|2|3\n4|5|NA", file = "file.txt")`

`read_csv`("file.csv") কমা ডিলিমিটার সহ ফাইল পড়ুন।
`write_file("A,B,C\n1,2,3\n4,5,NA", file = "file.csv")`

`read_csv2`("file2.csv") সেমিকোলন ডিলিমিটার সহ ফাইল পড়ুন।
`write_file("A;B;C\n1,5;2;3\n4,5;5;NA", file = "file2.csv")`

`read_tsv`("file.tsv") ট্যাব ডিলিমিটার সহ ফাইল পড়ুন। এছাড়াও `read_table()`, `read_fwf`("file.tsv", fwf_widths(c(2, 2, NA))) সমান প্রস্থ সহ ফাইল পড়ুন।
`write_file("\tB\tC\n1\t2\t3\n4\t5\tNA\n", file = "file.tsv")`

যেকোনো প্রকল্পের অন্যতম প্রথম ধাপ হলে বাইরের ডেটা R-এ আমদানি করা। ডেটা বেশিরভাগ সময় ট্যাবুলার বিন্যাসে থাকে। যেমনঃ csv ফাইল বা স্প্রেডশিট

অন্যান্য প্রকারের ডেটা নিম্নোক্ত প্যাকেজগুলো ব্যবহার করুন :

- haven - SPSS, Stata, এবং SAS ফাইল
- DBI - ডেটাবেজ
- jsonlite - json
- xml2 - XML
- httr - Web APIs
- rvest - HTML (ওয়েব স্ক্রাপিং)
- readr::read_lines() - text ডেটা

এই শিট এর প্রথম পৃষ্ঠায় থাকছে কিভাবে readr এর মাধ্যমে text ফাইল পড়া এবং লেখা যায়।

পরের পৃষ্ঠায় থাকছে কিভাবে readxl দিয়ে Excel ফাইল পড়া যায় এবং googlesheets4 দিয়ে Google Sheets পড়া যায়।

readr এর মাধ্যমে কলাম নির্দিষ্টকরণ

কলাম নির্দিষ্টকরণ এর মাধ্যমে ঠিক করা হয় আমদানিকৃত কলামগুলোর ডেটা এর প্রকার কি হবে। গতানুগতিক ভাবে readr একটি কলাম নির্দিষ্টকরণ সৃষ্টি করবে এবং সারসংক্ষেপ আউটপুট আকারে দেখাবে।

`spec(x)` আমদানিকৃত data frame এর সম্পূর্ণ কলাম নির্দিষ্টকরণ বের করুন।

```
spec(x)
# cols(
#   age = col_integer(), age হচ্ছে integer
#   sex = col_character(), sex হচ্ছে character
#   earn = col_double()
# )
earn হচ্ছে double (numeric)
```

দরকারি কলাম আর্গুমেন্ট কলাম নির্দিষ্টকরণ বার্তা লুকান।
`read_*(file, show_col_types = FALSE)`

আমদানি করার জন্য কলাম নির্বাচন করুন নাম, অবস্থান, অথবা selection helpers ব্যবহার করুন।
`read_*(file, col_select = c(age, earn))`

কলাম এর প্রকার অনুমান করুন। কলাম এর প্রকার অনুমান করতে `read_*`() প্রথম 1000 সারির ডেটা দেখে। `guess_max` ব্যবহার করে এটা বাড়তে পারেন।
`read_*(file, guess_max = Inf)`

কলাম নির্দিষ্টকরণ নির্ধারণ গতানুগতিক প্রকার নির্ধারণ করুন।
`read_csv(
 file,
 col_type = list(default = col_double())
)`

কলাম প্রকার অথবা string সংক্ষেপ ব্যবহার করুন।
`read_csv(
 file,
 col_type = list(x = col_double(), y = "l", z = "_")
)`

একক string সংক্ষেপ ব্যবহার করুন।
col types: skip, guess, integer, logical, character
`read_csv(
 file,
 col_type = "_?ilc"
)`

কলামের প্রকার প্রত্যেক কলাম প্রকারের একটা ফাংশন এবং string সংক্ষেপ আছে।

- `col_logical()` - "l"
- `col_integer()` - "i"
- `col_double()` - "d"
- `col_number()` - "n"
- `col_character()` - "c"
- `col_factor(levels, ordered = FALSE)` - "f"
- `col_datetime(format = "")` - "T"
- `col_date(format = "")` - "D"
- `col_time(format = "")` - "t"
- `col_skip()` - "-", "_"
- `col_guess()` - "?"

দরকারি ফাইল পড়ার আর্গুমেন্ট

কলাম শিরোনাম ছাড়া `read_csv`("file.csv", col_names = FALSE)

শিরোনাম সহ `read_csv`("file.csv", col_names = c("x", "y", "z"))

একাধিক ফাইল এক টেবিলে পড়ুন।
`read_csv(c("f1.csv", "f2.csv", "f3.csv"), id = "origin_file")`

লাইন এড়িয়ে যান।
`read_csv("file.csv", skip = 1)`

কিছু লাইন পড়ুন।
`read_csv("file.csv", n_max = 1)`

নিরুদ্দিষ্ট হলে মান উল্লেখ করুন।
`read_csv("file.csv", na = c("1"))`

দশমিক চিহ্ন উল্লেখ করুন।
`read_delim("file2.csv", locale = locale(decimal_mark = ","))`

readr দিয়ে ডাটা সংরক্ষণ করুন

`write_*`(x, file, na = "NA", append, col_names, quote, escape, eol, num_threads, progress)

`write_delim`(x, file, delim = " ") যেকোনো ডিলিমিটার দিয়ে ফাইল লিখুন।
`write_csv`(x, file) কমা ডিলিমিটার দিয়ে ফাইল লিখুন।
`write_csv2`(x, file) সেমিকোলন ডিলিমিটার দিয়ে ফাইল লিখুন।
`write_tsv`(x, file) ট্যাব ডিলিমিটার দিয়ে ফাইল লিখুন।



স্প্রেডশিট আমদানি

readxl এর মাধ্যমে

EXCEL ফাইল পড়ুন



	A	B	C	D	E
1	x1	x2	x3	x4	x5
2	x		z	8	
3	y	7		9	10

s1

`read_excel(path, sheet = NULL, range = NULL)` .xls অথবা .xlsx ফাইল পড়ুন ফাইল extension এর উপর ভিত্তি করে। আরো `read` আর্গুমেন্ট এর জন্য প্রথম পৃষ্ঠা দেখুন। এছাড়া `read_xls()` এবং `read_xlsx()`।
`read_excel("excel_file.xlsx")`

শিট পড়ুন

A	B	C	D	E

s1 s2 s3

`read_excel(path, sheet = NULL)` কোন শিট পড়তে হবে তা অবস্থান অথবা নাম দিয়ে নির্ধারণ করুন।
`read_excel(path, sheet = 1)`
`read_excel(path, sheet = "s1")`

s1	s2	s3

`excel_sheets(path)` শিটগুলোর নামের vector পান
`excel_sheets("excel_file.xlsx")`

একাধিক শিট পড়ুন:

1. File path ব্যবহার করে শিটগুলোর নামের vector পান
2. vector নামগুলো শিট এর নাম হিসেবে নির্ধারণ করুন
3. `purrr::map_dfr()` ব্যবহার করে একাধিক ফাইল এক data frame এ আমদানি করুন

```
path <- "your_file_path.xlsx"
path %>% excel_sheets() %>%
  set_names() %>%
  map_dfr(read_excel, path = path)
```

অন্যান্য দরকারি EXCEL প্যাকেজ

Excel ফাইল এ ডেটা লেখার জন্য দেখুন:

- `openxlsx`
- `writexl`

নন-টাবুলার Excel ডেটা এর জন্য দেখুন:

- `tidyxl`

READXL এর কলাম নির্দিষ্টকরণ

কলাম নির্দিষ্টকরণ এর মাধ্যমে ঠিক করা হয় আমদানিকৃত কলামগুলোর ডেটা এর প্রকার কি হবে।
`read_excel()` এর `col_types` আর্গুমেন্ট ব্যবহার করে কলাম নির্দিষ্টকরণ নির্ধারণ করুন।

কলাম এর প্রকার অনুমান করুন
 কলান এর প্রকার অনুমান করতে `read_excel()` প্রথম 1000 সারির ডেটা দেখে। `guess_max` ব্যবহার করে এটা বাড়াতে পারেন।

`read_excel(path, guess_max = Inf)`

সকল কলামকে একই প্রকারের করুন।

যেমন: `character`
`read_excel(path, col_types = "text")`

প্রত্যেক কলাম আলাদা ভাবে নির্ধারণ করুন

`read_excel(
 path,
 col_types = c("text", "guess", "guess", "numeric")
)`

কলাম এর প্রকার

logical	numeric	text	date	list
TRUE	2	hello	1947-01-08	hello
FALSE	3.45	world	1956-10-21	1

- skip
- guess
- logical
- numeric
- text
- date
- list

যেসব কলামে একাধিক প্রকারের ডেটা আছে, সেখানে list ব্যবহার করুন।
 list-column ডেটা এর জন্য `tidyr` এবং `purrr` দেখুন।

READXL এবং GOOGLESHEETS4 এর জন্য কলাম নির্দিষ্টকরণ

A	B	C	D	E
1	1	2	3	4
2	x		y	z
3	6	7		9

s1

`readxl::read_excel()` অথবা `googlesheets4::read_sheet()` এর `range` আর্গুমেন্ট ব্যবহার করে একটি শিট এর cell গুলোর উপসেট পড়ুন
`read_excel(path, range = "Sheet1!B1:D2")`
`read_sheet(ss, range = "B1:D2")`

এছাড়া `range` আর্গুমেন্ট কলাম নির্দিষ্টকরণ ফাংশন এর সাথে ব্যবহার করুন
`cell_limits()`, `cell_rows()`, `cell_cols()`, and `anchored()`।

googlesheets4 এর মাধ্যমে

শিট পড়ুন

A	B	C	D	E
1	x1	x2	x3	x4
2	x		z	8
3	y	7		9

s1

`read_sheet(ss, sheet = NULL, range = NULL)` URL, শিট ID অথবা `googledrive` প্যাকেজ এর `dribble` থেকে শিট পড়ুন। প্রথম পৃষ্ঠা দেখুন আরও `read` আর্গুমেন্ট এর জন্য। `range_read()` এর মতই।

শিট এর METADATA

URL এর গঠন:
`https://docs.google.com/spreadsheets/d/SPREADSHEET_ID/edit#gid=SHEET_ID`

`gs4_get(ss)` স্প্রেডশিট এর meta data পান

`gs4_find(...)` সব স্প্রেডশিট ফাইল গুলোর ডেটা পান

`sheet_properties(ss)` প্রত্যেক worksheet এর বৈশিষ্ট্য সম্বলিত tibble পান।
 এছাড়া `sheet_names()`

শিটে লিখুন

1	x	4
2	y	5
3	z	6

A	B	C	D
1			
2			

s1

x1	x2	x3
1	x1	x2
2	y	5
3	z	6

s1

`write_sheet(data, ss = NULL, sheet = NULL)` নতুন অথবা বিদ্যমান শিটে data frame লিখুন

`gs4_create(name, ..., sheets = NULL)` নামের vector, data frame, অথবা a (নাম সম্বলিত) data frame এর list দিয়ে নতুন শিট বানান

`sheet_append(ss, data, sheet = 1)` worksheet এর শেষে একটি সারি যুক্ত করুন

GOOGLESHEETS4 এর কলাম নির্দিষ্টকরণ

কলাম নির্দিষ্টকরণ এর মাধ্যমে ঠিক করা হয় আমদানিকৃত কলামগুলোর ডেটা এর প্রকার কি হবে।

`read_sheet()/range_read()` এর `col_types` আর্গুমেন্ট ব্যবহার করে কলাম নির্দিষ্টকরণ নির্ধারণ করুন।

কলাম এর প্রকার অনুমান করুন
 কলান এর প্রকার অনুমান করতে

`read_sheet()/range_read()` প্রথম 1000 সারির ডেটা দেখে। `guess_max` ব্যবহার করে এটা বাড়াতে পারেন।
`read_sheet(path, guess_max = Inf)`

সকল কলামকে একই প্রকারের করুন।

যেমন: `character`
`read_sheet(path, col_types = "c")`

প্রত্যেক কলাম আলাদা ভাবে নির্ধারণ করুন

col types: skip, guess, integer, logical, character
`read_sheets(ss, col_types = "?_?ilc")`

কলাম এর প্রকার

l	n	c	D	L
TRUE	2	hello	1947-01-08	hello
FALSE	3.45	world	1956-10-21	1

- skip - "_" or "-"
- guess - "?"
- logical - "l"
- integer - "i"
- double - "d"
- numeric - "n"
- date - "D"
- datetime - "T"
- character - "c"
- list-column - "L"
- cell - "C" Returns list of raw cell data.

যেসব কলামে একাধিক প্রকারের ডেটা আছে, সেখানে list ব্যবহার করুন।
 list-column ডেটা এর জন্য `tidyr` এবং `purrr` দেখুন।

ফাইল স্তরীয় অপারেশন

এছাড়াও `googlesheets4` শিটের বিভিন্ন দিক পরিবর্তন করার উপায় প্রদান করে। (যেমন: সারি freeze করা, কলামের প্রস্থ নির্ধারণ করা, (work)sheets বাবস্থাপনা করা)। আরও জানতে [googlesheets4.tidyverse.org](https://tidyverse.org/guides/googlesheets4/) এ যান।

সম্পূর্ণ ফাইল অপারেশনের (যেমন: rename করা, share করা, folder এ রাখা) জন্য `tidyverse` এর অন্তর্ভুক্ত প্যাকেজ `googledrive` দেখুন
[googledrive.tidyverse.org](https://tidyverse.org/guides/googledrive/).

