R Gather Data Columns from Multiple CSV Files

Fan Wang

2021-01-10

Contents

1 Gather Files

Go to the **RMD**, **R**, **PDF**, or **HTML** version of this file. Go back to fan's REconTools Package, R Code Examples Repository (bookdown site), or Intro Stats with R Repository (bookdown site).

1.1 Stack CSV Files Together Extract and Select Variables

There are multiple csv files, each was simulated with a different combination of parameters, each file has the same columns and perhaps even the same number of rows. We want to combine the files together, and provide correct attributes to rows from each table stacked, based on each underlying csv file's file name.

This is necessary, for example, when running computational exercises across EC2 instances in batch array and files are saved to different S3 folders. Need to gather parallel computational results together in a single file after syncing files locally with S3.

In the csv folder under this section, there are four subfolder, each containing 3 files with identical file structures.

We want to find the relevant csv files from these directories, and stack the results together.

1. File search search string, search in all subfolders, the search string contains file prefix that is common across files that need to be gathered.
2. Extract path folder hierarchy, each layer of folder is a different variable
3. Stack files together, with variables for file name and folder name
4. Extract from file name the component that is not in the search string, keep as separate variable
5. Follow specific rules about how file suffix is constructed to obtain additional variables.
6. Keep only a subset of columns of interest.

First, search and find all files with certain prefix.

```
# can search in multiple paths, second path here has no relevant contents
spt_roots <- c('C:/Users/fan/R4Econ/panel/basic/_file/csv',
               'C:/Users/fan/R4Econ/panel/basic/_file/tex')
# can skip file names with certain strings
spn_skip <- c('A3420')
# prefix search patther,
st_search_str <- 'solu_19E1NEp99r99x_ITG_PE_cev_*'

# Search and get all Path
ls_sfls <- list.files(path=spt_roots,
```

```
                      recursive=T,
                      pattern=st_search_str,
                      full.names=T)

# Skip path if contains words in skip list
if(!missing(spn_skip)) {
  ls_sfls <- ls_sfls[!grepl(paste(spn_skip, collapse = "|"), ls_sfls)]
}
```

Second, show all the files found, show their full path, the file name and the two folder names above the file name.

```
# Loop and print found files
it_folders_names_to_keep = 2
for (spt_file in ls_sfls) {
    ls_srt_folders_name_keep <- tail(strsplit(spt_file, "/")[[1]], n=it_folders_names_to_keep+1)
    snm_file_name <- tail(ls_srt_folders_name_keep, 1)
    ls_srt_folders_keep <- head(ls_srt_folders_name_keep, it_folders_names_to_keep)
    print(paste0('path:', spt_file))
    print(snm_file_name)
    print(ls_srt_folders_keep)
}
```

```
## [1] "path:C:/Users/fan/R4Econ/panel/basic/_file/csv/cev-2000/solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-200
## [1] "solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A0.csv"
## [1] "csv"      "cev-2000"
## [1] "path:C:/Users/fan/R4Econ/panel/basic/_file/csv/cev-2000/solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-200
## [1] "solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A6840.csv"
## [1] "csv"      "cev-2000"
## [1] "path:C:/Users/fan/R4Econ/panel/basic/_file/csv/cev-947/solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_
## [1] "solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A0.csv"
## [1] "csv"     "cev-947"
## [1] "path:C:/Users/fan/R4Econ/panel/basic/_file/csv/cev-947/solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_
## [1] "solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A6840.csv"
## [1] "csv"     "cev-947"
## [1] "path:C:/Users/fan/R4Econ/panel/basic/_file/csv/cev2000/solu_19E1NEp99r99x_ITG_PE_cev_c19_cev200
## [1] "solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A0.csv"
## [1] "csv"     "cev2000"
## [1] "path:C:/Users/fan/R4Econ/panel/basic/_file/csv/cev2000/solu_19E1NEp99r99x_ITG_PE_cev_c19_cev200
## [1] "solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A6840.csv"
## [1] "csv"     "cev2000"
## [1] "path:C:/Users/fan/R4Econ/panel/basic/_file/csv/cev947/solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947_A
## [1] "solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947_A0.csv"
## [1] "csv"    "cev947"
## [1] "path:C:/Users/fan/R4Econ/panel/basic/_file/csv/cev947/solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947_A
## [1] "solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947_A6840.csv"
## [1] "csv"    "cev947"
```

Third, create a dataframe with the folder and file names:

```
# String matrix empty
mt_st_paths_names <- matrix(data=NA, nrow=length(ls_sfls), ncol=4)

# Loop and print found files
it_folders_names_to_keep = 2
it_file_counter = 0
```

```
for (spt_file in ls_sfls) {
    # row counter
    it_file_counter = it_file_counter + 1

    # get file paths
    ls_srt_folders_name_keep <- tail(strsplit(spt_file, "/")[[1]], n=it_folders_names_to_keep+1)
    snm_file_name <- tail(ls_srt_folders_name_keep, 1)
    ls_srt_folders_keep <- head(ls_srt_folders_name_keep, it_folders_names_to_keep)

    # store
    # tools::file_path_sans_ext to drop suffix
    mt_st_paths_names[it_file_counter,1] = tools::file_path_sans_ext(snm_file_name)
    mt_st_paths_names[it_file_counter,2] = ls_srt_folders_keep[1]
    mt_st_paths_names[it_file_counter,3] = ls_srt_folders_keep[2]
    mt_st_paths_names[it_file_counter,4] = spt_file
}

# Column Names
ar_st_varnames <- c('fileid','name','folder1','folder2', 'fullpath')

# Combine to tibble, add name col1, col2, etc.
tb_csv_info <- as_tibble(mt_st_paths_names) %>%
  rowid_to_column(var = "id") %>%
  rename_all(~c(ar_st_varnames))

# Display
kable(tb_csv_info[,1:4]) %>% kable_styling_fc()
```

fileid	name	folder1	folder2
1	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A0	csv	cev-2000
2	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A6840	csv	cev-2000
3	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A0	csv	cev-947
4	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A6840	csv	cev-947
5	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A0	csv	cev2000
6	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A6840	csv	cev2000
7	solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947_A0	csv	cev947
8	solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947_A6840	csv	cev947

Fourth, create a dataframe by expanding each row with the datafile loaded in, use apply with anonymous function.

```
# Generate a list of dataframes
ls_df_loaded_files =
  apply(tb_csv_info,
        1,
        function(row) {
          # Loading file
          spn_full_path <- row[5]
          mt_csv = read.csv(file = spn_full_path)
          # dataframe
          it_fileid <- row[1]
          snm_filename <- row[2]
          srt_folder_level2 <- row[3]
          srt_folder_level1 <- row[4]
```

```
          tb_combine = as_tibble(mt_csv) %>%
            na.omit %>%
            rowid_to_column(var = "statesid") %>%
            mutate(fileid = it_fileid,
                   filename = snm_filename,
                   folder_lvl1 = srt_folder_level1,
                   folder_lvl2 = srt_folder_level2) %>%
            select(fileid, filename, folder_lvl1, folder_lvl2,
                   statesid, everything())
          # return
          return(tb_combine)
        })

# Stack dataframes together
df_all_files = do.call(bind_rows, ls_df_loaded_files)

# show stacked table
kable(df_all_files[seq(1,601,50),1:6]) %>% kable_styling_fc_wide()
```

fileid	filename	folder_lvl1	folder_lvl2	statesid	EjV
1	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A0	cev-2000	csv	1	-28.8586860
1	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A0	cev-2000	csv	51	-0.2106603
2	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A6840	cev-2000	csv	3	-28.8586860
2	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A6840	cev-2000	csv	53	-0.0642997
3	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A0	cev-947	csv	5	-5.8826609
3	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A0	cev-947	csv	55	0.0353187
4	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A6840	cev-947	csv	7	-2.7046907
4	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A6840	cev-947	csv	57	0.1094474
5	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A0	cev2000	csv	9	-2.9782236
5	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A0	cev2000	csv	59	0.3389275
6	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A6840	cev2000	csv	11	-1.7229647
6	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A6840	cev2000	csv	61	-14.6880377
7	solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947_A0	cev947	csv	13	-1.7623279

Fifth, get additional information from the file name and file folder. Extract those as separate variables. The file names is dash connected, with various information. First, split just the final element of the string file name out, which is *A###*. Then, also extract the number next to N as a separate numeric column. Additional *folder_lvl1* separate out the numeric number from the initial word *cev*.

Split "solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A###" to "solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000" and "A###":

```
# separate last eleemtnafter underscore
df_all_files_finalA <- df_all_files %>%
   separate(filename, into = c("filename_main", "prod_type_st"),
            sep="_(?=[^_]+$)",
            remove = FALSE) %>%
   select(fileid, filename, filename_main, prod_type_st, folder_lvl1, folder_lvl2,
          statesid, everything())
# show stacked table
kable(df_all_files_finalA[seq(1,601,50),1:10]) %>% kable_styling_fc_wide()
```

Split "A###" to "A" and "A###". Additionally, also split *cev####* to *cev* and *####*, allow for positive and negative numbers. See regular expression 101 helper

fileid	filename	filename_main	prod_type_st	folder_lvl1	folder_lvl2	statesid	EjV	k_tt	b_tt
1	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A0	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000	A0	cev-2000	csv	1	-28.8586860	0.000000	0.000000
1	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A0	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000	A0	cev-2000	csv	51	-0.2106603	0.000000	78.005300
2	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A6840	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000	A6840	cev-2000	csv	3	-28.8586860	0.000000	0.000000
2	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000_A6840	solu_19E1NEp99r99x_ITG_PE_cev_c0_cev-2000	A6840	cev-2000	csv	53	-0.0642997	0.000000	84.215368
3	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A0	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947	A0	cev-947	csv	5	-5.8826609	0.000000	3.869909
3	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A0	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947	A0	cev-947	csv	55	0.0353187	0.000000	90.611739
4	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A6840	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947	A6840	cev-947	csv	7	-2.7046907	0.000000	7.855916
4	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947_A6840	solu_19E1NEp99r99x_ITG_PE_cev_c5_cev-947	A6840	cev-947	csv	57	0.1094474	0.000000	90.611739
5	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A0	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000	A0	cev2000	csv	9	-2.9782236	0.000000	7.855916
5	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A0	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000	A0	cev2000	csv	59	0.3389275	0.000000	97.200000
6	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A6840	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000	A6840	cev2000	csv	11	-1.7229647	0.000000	11.961502
6	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000_A6840	solu_19E1NEp99r99x_ITG_PE_cev_c19_cev2000	A6840	cev2000	csv	61	-14.6880377	1.990694	-1.879215
7	solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947_A0	solu_19E1NEp99r99x_ITG_PE_cev_c14_cev947	A0	cev947	csv	13	-1.7623279	0.000000	16.190257

```
# string and number separation
df_all_files_finalB <- df_all_files_finalA %>%
  separate(prod_type_st,
           into = c("prod_type_st_prefix", "prod_type_lvl"),
           sep="(?<=[A-Za-z])(?=[-0-9])", # positive or negative numbers
           remove=FALSE) %>%
  separate(folder_lvl1,
           into = c("cev_prefix", "cev_lvl"),
           sep="(?<=[A-Za-z])(?=[-0-9])", # positive or negative numbers
           remove=FALSE) %>%
  mutate(cev_st = folder_lvl1,
         prod_type_lvl = as.numeric(prod_type_lvl),
         cev_lvl = as.numeric(cev_lvl)/10000) %>%
  select(fileid,
         prod_type_st, prod_type_lvl,
         cev_st, cev_lvl,
         statesid, EjV,
         filename, folder_lvl1, folder_lvl2)
# Ordering, sort by cev_lvl, then prod_type_lvl, then stateid
df_all_files_finalB <- df_all_files_finalB %>%
  arrange(cev_lvl, prod_type_lvl, statesid)
# show stacked table
kable(df_all_files_finalB[seq(1,49*16,49),1:7]) %>% kable_styling_fc_wide()
```

fileid	prod_type_st	prod_type_lvl	cev_st	cev_lvl	statesid	EjV
1	A0	0	cev-2000	-0.2000	1	-28.8586860
1	A0	0	cev-2000	-0.2000	50	-0.2106603
2	A6840	6840	cev-2000	-0.2000	1	-28.8586860
2	A6840	6840	cev-2000	-0.2000	50	-0.1311749
3	A0	0	cev-947	-0.0947	1	-28.0399281
3	A0	0	cev-947	-0.0947	50	-0.0911499
4	A6840	6840	cev-947	-0.0947	1	-28.0399281
4	A6840	6840	cev-947	-0.0947	50	-0.0134719
7	A0	0	cev947	0.0947	1	-26.8243673
7	A0	0	cev947	0.0947	50	0.0857474
8	A6840	6840	cev947	0.0947	1	-26.8243673
8	A6840	6840	cev947	0.0947	50	0.1608382
5	A0	0	cev2000	0.2000	1	-26.2512036
5	A0	0	cev2000	0.2000	50	0.1694524
6	A6840	6840	cev2000	0.2000	1	-26.2512036
6	A6840	6840	cev2000	0.2000	50	0.2432677