0_5_tidyverse_group

March 17, 2023

1 Einführung in tidyverse - Teil 2

Es geht weiter mit tidyverse.

Wir arbeiten mit dem Datensatz ae weiter, den wir schon kennen.

Wir decken in dieser Datei die SAS-Konzepte

- PROC SORT
- FIRST / LAST
- BY

ab.

1.1 Daten laden

Nur, wenn das Programm zwischenzeitlich beendet worden ist.

[1]:
```
library(tidyverse)
```

```
  Attaching packages

tidyverse 1.3.2
 ggplot2 3.4.0       purrr   0.3.5
 tibble  3.1.8       dplyr   1.0.10
 tidyr   1.2.1       stringr 1.4.1
 readr   2.1.3       forcats 0.5.2
  Conflicts
            tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()
```

[2]:
```
ae = read.csv(file = "data/ae.csv", sep = ",", header = TRUE, quote = "\"'")
```

[3]:
```
# Wir wollen sehen, wie viele unterschiedliche Patienten AEs hatten
ae %>% group_by(USUBJID) %>% count() %>% ungroup()
ae %>% group_by(USUBJID) %>% count() %>% ungroup() %>% nrow()
ae %>% nrow()
```

USUBJID	n
<chr>	<int>
01-701-1015	3
01-701-1023	4
01-701-1028	2
01-701-1034	2
01-701-1047	4
01-701-1097	10
01-701-1111	8
01-701-1115	9
01-701-1118	1
01-701-1130	8
01-701-1133	4
01-701-1146	11
01-701-1148	10
01-701-1153	2
01-701-1180	9
01-701-1181	1
01-701-1188	8
01-701-1192	15
01-701-1203	1
01-701-1211	9
01-701-1239	10
01-701-1275	15
01-701-1287	5
01-701-1294	6
01-701-1302	23
01-701-1317	9
01-701-1324	4
01-701-1341	5
01-701-1360	3
A tibble: 225 × 2 01-701-1363	6
...	...
01-716-1177	1
01-716-1189	4
01-716-1229	2
01-716-1298	6
01-716-1308	1
01-716-1311	4
01-716-1364	2
01-716-1373	2
01-716-1418	10
01-716-1441	1
01-716-1447	5
01-717-1004	19
01-717-1109	8
01-717-1174	6
01-717-1201	2
01-717-1344	5
01-717-1357	9
01-717-1446	9
01-718-1066	4
01-718-1079	3

225

1191

1.2 Daten sortieren

[4]:
```
# Aufsteigend nach Zahl der AEs sortieren
# Sortieren nach mehreren Variablen, im Befehl **arrange()** in gewünschter␣
 ↪Reihenfolge auflisten.
ae %>% arrange(AESTDY) %>% head() %>% select(USUBJID, AETERM, AESTDY)
#ae %>% filter(USUBJID == "01-705-1393")
```

A data.frame: 6 × 3

	USUBJID <chr>	AETERM <chr>	AESTDY <int>
1	01-705-1393	PRURITUS	-277
2	01-705-1393	PRURITUS	-277
3	01-711-1433	HYPERTENSION	-188
4	01-711-1433	HYPERTENSION	-188
5	01-704-1388	HEADACHE	-106
6	01-701-1111	LOCALISED INFECTION	-61

[5]:
```
# Gruppieren und sortieren
ae %>% group_by(USUBJID) %>% count() %>% ungroup() %>% arrange(n)
```

USUBJID	n
<chr>	<int>
01-701-1118	1
01-701-1181	1
01-701-1203	1
01-701-1442	1
01-703-1175	1
01-703-1295	1
01-704-1218	1
01-704-1388	1
01-704-1435	1
01-704-1445	1
01-705-1031	1
01-705-1059	1
01-705-1186	1
01-705-1280	1
01-708-1032	1
01-708-1372	1
01-709-1301	1
01-709-1424	1
01-710-1083	1
01-710-1187	1
01-713-1073	1
01-714-1425	1
01-715-1207	1
01-715-1319	1
01-716-1177	1
01-716-1308	1
01-716-1441	1
01-701-1028	2
01-701-1034	2
01-701-1153	2
...	...
01-717-1357	9
01-717-1446	9
01-718-1254	9
01-701-1097	10
01-701-1148	10
01-701-1239	10
01-702-1082	10
01-704-1065	10
01-706-1384	10
01-716-1418	10
01-718-1150	10
01-701-1146	11
01-709-1217	11
01-709-1259	11
01-718-1250	11
01-701-1383	12
01-708-1272	12
01-710-1006	12
01-710-1045	13
01-718-1355	13

A tibble: 225 × 2

[6]:
```
# Absteigend nach Zahl der AEs sortieren
ae %>% group_by(USUBJID) %>% count() %>% ungroup() %>% arrange(desc(n)) %>%␣
 ↪head(10)
```

A tibble: 10 × 2

USUBJID <chr>	n <int>
01-701-1302	23
01-717-1004	19
01-704-1266	16
01-709-1029	16
01-718-1427	16
01-701-1192	15
01-701-1275	15
01-709-1309	15
01-713-1179	15
01-711-1143	14

1.3 Funktionen zum Aggregieren

Standardfunktionen

- min
- max
- mean
- median
- var
- sd

[7]:
```
# Neuen Datensatz SDTM DM laden
dm = read.csv(file = "data/dm.csv", sep = ",", header = TRUE, quote = "\"'")
head(dm)
glimpse(dm)
```

A data.frame: 6 × 25

	STUDYID <chr>	DOMAIN <chr>	USUBJID <chr>	SUBJID <int>	RFSTDTC <chr>	RFENDTC <chr>	RF <c
1	CDISCPILOT01	DM	01-701-1015	1015	2014-01-02	2014-07-02	201
2	CDISCPILOT01	DM	01-701-1023	1023	2012-08-05	2012-09-02	201
3	CDISCPILOT01	DM	01-701-1028	1028	2013-07-19	2014-01-14	201
4	CDISCPILOT01	DM	01-701-1033	1033	2014-03-18	2014-04-14	201
5	CDISCPILOT01	DM	01-701-1034	1034	2014-07-01	2014-12-30	201
6	CDISCPILOT01	DM	01-701-1047	1047	2013-02-12	2013-03-29	201

```
Rows: 306
Columns: 25
$ STUDYID  <chr> "CDISCPILOT01", "CDISCPILOT01",
"CDISCPILOT01", "CDISCPILOT01…
$ DOMAIN   <chr> "DM", "DM", "DM", "DM", "DM", "DM", "DM",
"DM", "DM", "DM", "…
$ USUBJID  <chr> "01-701-1015", "01-701-1023", "01-701-1028",
```

```
"01-701-1033", "…
$ SUBJID   <int> 1015, 1023, 1028, 1033, 1034, 1047, 1057,
1097, 1111, 1115, 1…
$ RFSTDTC  <chr> "2014-01-02", "2012-08-05", "2013-07-19",
"2014-03-18", "2014…
$ RFENDTC  <chr> "2014-07-02", "2012-09-02", "2014-01-14",
"2014-04-14", "2014…
$ RFXSTDTC <chr> "2014-01-02", "2012-08-05", "2013-07-19",
"2014-03-18", "2014…
$ RFXENDTC <chr> "2014-07-02", "2012-09-01", "2014-01-14",
"2014-03-31", "2014…
$ RFICDTC  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ RFPENDTC <chr> "2014-07-02T11:45", "2013-02-18",
"2014-01-14T11:10", "2014-0…
$ DTHDTC   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ DTHFL    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ SITEID   <int> 701, 701, 701, 701, 701, 701, 701, 701, 701,
701, 701, 701, 7…
$ AGE      <int> 63, 64, 71, 74, 77, 85, 59, 68, 81, 84, 52,
84, 81, 57, 75, 5…
$ AGEU     <chr> "YEARS", "YEARS", "YEARS", "YEARS", "YEARS",
"YEARS", "YEARS"…
$ SEX      <chr> "F", "M", "M", "M", "F", "F", "F", "M", "F",
"M", "M", "M", "…
$ RACE     <chr> "WHITE", "WHITE", "WHITE", "WHITE", "WHITE",
"WHITE", "WHITE"…
$ ETHNIC   <chr> "HISPANIC OR LATINO", "HISPANIC OR LATINO",
"NOT HISPANIC OR …
$ ARMCD    <chr> "Pbo", "Pbo", "Xan_Hi", "Xan_Lo", "Xan_Hi",
"Pbo", "Scrnfail"…
$ ARM      <chr> "Placebo", "Placebo", "Xanomeline High
Dose", "Xanomeline Low…
$ ACTARMCD <chr> "Pbo", "Pbo", "Xan_Hi", "Xan_Lo", "Xan_Hi",
"Pbo", "Scrnfail"…
$ ACTARM   <chr> "Placebo", "Placebo", "Xanomeline High
Dose", "Xanomeline Low…
$ COUNTRY  <chr> "USA", "USA", "USA", "USA", "USA", "USA",
"USA", "USA", "USA"…
$ DMDTC    <chr> "2013-12-26", "2012-07-22", "2013-07-11",
"2014-03-10", "2014…
$ DMDY     <int> -7, -14, -8, -8, -7, -21, NA, -9, -13, -7,
-13, -6, -5, NA, -…
```

[8]:
```
# Nacharbeiten des Import
dm = read_csv(file = "data/dm.csv")
```

```
Rows: 306 Columns: 25
  Column specification


Delimiter: ","
chr  (13): STUDYID, DOMAIN, USUBJID, DTHFL, AGEU, SEX, RACE, ETHNIC,
ARMCD, …
dbl   (4): SUBJID, SITEID, AGE, DMDY
lgl   (1): RFICDTC
dttm  (1): RFPENDTC
date  (6): RFSTDTC, RFENDTC, RFXSTDTC, RFXENDTC, DTHDTC, DMDTC

 Use `spec()` to retrieve the full column specification for this
data.
 Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

[9]:
```
head(dm)

table(dm$RFICDTC)
```

A tibble: 6 × 25

STUDYID <chr>	DOMAIN <chr>	USUBJID <chr>	SUBJID <dbl>	RFSTDTC <date>	RFENDTC <date>	RFXSTDT <date>
CDISCPILOT01	DM	01-701-1015	1015	2014-01-02	2014-07-02	2014-01-02
CDISCPILOT01	DM	01-701-1023	1023	2012-08-05	2012-09-02	2012-08-05
CDISCPILOT01	DM	01-701-1028	1028	2013-07-19	2014-01-14	2013-07-19
CDISCPILOT01	DM	01-701-1033	1033	2014-03-18	2014-04-14	2014-03-18
CDISCPILOT01	DM	01-701-1034	1034	2014-07-01	2014-12-30	2014-07-01
CDISCPILOT01	DM	01-701-1047	1047	2013-02-12	2013-03-29	2013-02-12

```
< table of extent 0 >
```

[10]:
```
dm = read_csv(file = "data/dm.csv",
              col_types = list(
              RFICDTC = col_date(format = "")
              ))
head(dm)
table(dm$RFICDTC)
```

A tibble: 6 × 25

STUDYID <chr>	DOMAIN <chr>	USUBJID <chr>	SUBJID <dbl>	RFSTDTC <date>	RFENDTC <date>	RFXSTDT <date>
CDISCPILOT01	DM	01-701-1015	1015	2014-01-02	2014-07-02	2014-01-02
CDISCPILOT01	DM	01-701-1023	1023	2012-08-05	2012-09-02	2012-08-05
CDISCPILOT01	DM	01-701-1028	1028	2013-07-19	2014-01-14	2013-07-19
CDISCPILOT01	DM	01-701-1033	1033	2014-03-18	2014-04-14	2014-03-18
CDISCPILOT01	DM	01-701-1034	1034	2014-07-01	2014-12-30	2014-07-01
CDISCPILOT01	DM	01-701-1047	1047	2013-02-12	2013-03-29	2013-02-12

```
< table of extent 0 >
```

[11]:
```
# Summary-Funktionen
# Mögliche "Fehlfunktionen" bei fehlenden Werten
dm %>% summarise(AGE_mean = mean(AGE))
dm %>% filter(is.na(AGE)) %>% count()
dm %>% summarise(AGE_mean = mean(AGE, na.rm = TRUE))
```

A tibble: 1 × 1

AGE_mean
<dbl>
75.08824

A spec_tbl_df: 1 × 1

n
<int>
0

A tibble: 1 × 1

AGE_mean
<dbl>
75.08824

[12]:
```
# Mehrere Summary-Funktionen möglich
ae %>% summarise(AESTDY_n = n(),
                 AESTDY_mean = mean(AESTDY),
                 AESTDY_sd = sd(AESTDY))

ae %>% summarise(AESTDY_n = n(),
                 AESTDY_mean = mean(AESTDY, na.rm = TRUE),
                 AESTDY_sd = sd(AESTDY, na.rm = TRUE))
```

A data.frame: 1 × 3

AESTDY_n	AESTDY_mean	AESTDY_sd
<int>	<dbl>	<dbl>
1191	NA	NA

A data.frame: 1 × 3

AESTDY_n	AESTDY_mean	AESTDY_sd
<int>	<dbl>	<dbl>
1191	45.82833	48.22689

[13]:
```
# Mehrere Summary-Funktionen möglich
# Zusätzliches Gruppieren nach Treatment
dm %>% group_by(ACTARM) %>%
  summarise(AGE_n = n(),
            AGE_mean = mean(AGE, na.rm = TRUE),
            AGE_sd = sd(AGE, na.rm = TRUE))
```

A tibble: 4 × 4

ACTARM	AGE_n	AGE_mean	AGE_sd
<chr>	<int>	<dbl>	<dbl>
Placebo	86	75.20930	8.590167
Screen Failure	52	75.09615	9.699928
Xanomeline High Dose	72	73.77778	7.943856
Xanomeline Low Dose	96	75.95833	8.113558

1.4 FIRST und LAST

FIRST und LAST lassen sich nachbauen oder mit entsprechenden Funktionen nutzen.

[14]:
```
# Welcher Patient hat als erstes welche Therapie bekommen?
dm1 <- dm %>% select(USUBJID, ACTARM, RFXSTDTC)
head(dm1)
```

A tibble: 6 × 3

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>
01-701-1015	Placebo	2014-01-02
01-701-1023	Placebo	2012-08-05
01-701-1028	Xanomeline High Dose	2013-07-19
01-701-1033	Xanomeline Low Dose	2014-03-18
01-701-1034	Xanomeline High Dose	2014-07-01
01-701-1047	Placebo	2013-02-12

[15]:
```
# Sortieren nach ACTARM und RFXSTDTC
dm1 %>% arrange(ACTARM, RFXSTDTC) %>% head()
```

A tibble: 6 × 3

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>
01-716-1024	Placebo	2012-07-09
01-711-1036	Placebo	2012-07-29
01-701-1023	Placebo	2012-08-05
01-704-1260	Placebo	2012-08-30
01-703-1299	Placebo	2012-09-12
01-704-1164	Placebo	2012-09-19

[16]:
```
# Nutzen der first()- und last()-Funktionen
dm1 %>% arrange(ACTARM, RFXSTDTC) %>% group_by(ACTARM) %>%
  summarise(trt_first = first(USUBJID),
            trt_last = last(USUBJID)) %>%
  ungroup()
```

A tibble: 4 × 3

ACTARM <chr>	trt_first <chr>	trt_last <chr>
Placebo	01-716-1024	01-716-1177
Screen Failure	01-701-1057	01-716-1331
Xanomeline High Dose	01-703-1258	01-701-1034
Xanomeline Low Dose	01-701-1192	01-701-1317

[17]:
```
head(dm1)
```

A tibble: 6 × 3

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>
01-701-1015	Placebo	2014-01-02
01-701-1023	Placebo	2012-08-05
01-701-1028	Xanomeline High Dose	2013-07-19
01-701-1033	Xanomeline Low Dose	2014-03-18
01-701-1034	Xanomeline High Dose	2014-07-01
01-701-1047	Placebo	2013-02-12

[18]:
```
# Alternativer Ansatz unter Beibehaltung aller Daten und
# anschließendem Filtern
dm2 <- dm1 %>% arrange(ACTARM, RFXSTDTC) %>%
          group_by(ACTARM) %>%
          mutate(id = row_number()) %>%
          mutate(id_min = min(id)) %>%
          mutate(id_max = max(id)) %>%
          ungroup()
head(dm2)
```

A tibble: 6 × 6

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>	id <int>	id_min <int>	id_max <int>
01-716-1024	Placebo	2012-07-09	1	1	86
01-711-1036	Placebo	2012-07-29	2	1	86
01-701-1023	Placebo	2012-08-05	3	1	86
01-704-1260	Placebo	2012-08-30	4	1	86
01-703-1299	Placebo	2012-09-12	5	1	86
01-704-1164	Placebo	2012-09-19	6	1	86

[19]:
```
# Filtern kann auch nur für eine Bedingung erfolgen, um first und last getrennt␣
↪zu erhalten.
dm2 %>% filter(id == id_min | id == id_max)
```

A tibble: 8 × 6

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>	id <int>	id_min <int>	id_max <int>
01-716-1024	Placebo	2012-07-09	1	1	86
01-716-1177	Placebo	2014-09-02	86	1	86
01-701-1057	Screen Failure	NA	1	1	52
01-716-1331	Screen Failure	NA	52	1	52
01-703-1258	Xanomeline High Dose	2012-07-20	1	1	72
01-701-1034	Xanomeline High Dose	2014-07-01	72	1	72
01-701-1192	Xanomeline Low Dose	2012-07-22	1	1	96
01-701-1317	Xanomeline Low Dose	2014-05-22	96	1	96

1.5 Neue Variablen erzeugen

Wir haben oben den Befehl mutate() gesehen. Dieser erzeugt im Data Frame eine neue Variable. Hier können auch mehrere Variablen miteinander verknüpft werden.

[20]:
```
dm3 <- dm %>% select(USUBJID, ACTARM, RFXSTDTC, RFXENDTC)
head(dm3)
```

A tibble: 6 × 4

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>	RFXENDTC <date>
01-701-1015	Placebo	2014-01-02	2014-07-02
01-701-1023	Placebo	2012-08-05	2012-09-01
01-701-1028	Xanomeline High Dose	2013-07-19	2014-01-14
01-701-1033	Xanomeline Low Dose	2014-03-18	2014-03-31
01-701-1034	Xanomeline High Dose	2014-07-01	2014-12-30
01-701-1047	Placebo	2013-02-12	2013-03-09

[21]:
```
# Wir erhalten ein Datendifferenz-Objekt.
dm3 %>% mutate(dd = RFXENDTC - RFXSTDTC) %>% head()
```

A tibble: 6 × 5

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>	RFXENDTC <date>	dd <drtn>
01-701-1015	Placebo	2014-01-02	2014-07-02	181 days
01-701-1023	Placebo	2012-08-05	2012-09-01	27 days
01-701-1028	Xanomeline High Dose	2013-07-19	2014-01-14	179 days
01-701-1033	Xanomeline Low Dose	2014-03-18	2014-03-31	13 days
01-701-1034	Xanomeline High Dose	2014-07-01	2014-12-30	182 days
01-701-1047	Placebo	2013-02-12	2013-03-09	25 days

[22]:
```
# Jetzt gibt es einen Integerwert.
dm3 %>% mutate(dd = as.integer(RFXENDTC - RFXSTDTC)) %>% head()
```

A tibble: 6 × 5

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>	RFXENDTC <date>	dd <int>
01-701-1015	Placebo	2014-01-02	2014-07-02	181
01-701-1023	Placebo	2012-08-05	2012-09-01	27
01-701-1028	Xanomeline High Dose	2013-07-19	2014-01-14	179
01-701-1033	Xanomeline Low Dose	2014-03-18	2014-03-31	13
01-701-1034	Xanomeline High Dose	2014-07-01	2014-12-30	182
01-701-1047	Placebo	2013-02-12	2013-03-09	25

[23]:
```
# Alternative Zuweisung über $ möglich
# Geschmackssache ...
dm3$dd <- as.integer(dm3$RFXENDTC - dm3$RFXSTDTC)
head(dm3)
```

A tibble: 6 × 5

USUBJID <chr>	ACTARM <chr>	RFXSTDTC <date>	RFXENDTC <date>	dd <int>
01-701-1015	Placebo	2014-01-02	2014-07-02	181
01-701-1023	Placebo	2012-08-05	2012-09-01	27
01-701-1028	Xanomeline High Dose	2013-07-19	2014-01-14	179
01-701-1033	Xanomeline Low Dose	2014-03-18	2014-03-31	13
01-701-1034	Xanomeline High Dose	2014-07-01	2014-12-30	182
01-701-1047	Placebo	2013-02-12	2013-03-09	25

[ ]: