0_4_tidyverse

March 17, 2023

1 Einführung in tidyverse - Teil 1

tidyverse ist eine Sammlung von Paketen, die die Datenmanipulation standardisieren und erleichtern.

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

Wir decken in dieser Datei die SAS-Konzepte

- DROP / KEEP
- WHERE
- MERGE

ab.

1.1 Daten laden

[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]:
```
head(ae)
```

A data.frame: 6 × 35

	STUDYID <chr>	DOMAIN <chr>	USUBJID <chr>	AESEQ <int>	AESPID <chr>	AETERM <chr>
1	CDISCPILOT01	AE	01-701-1015	1	E07	APPLICATION SI
2	CDISCPILOT01	AE	01-701-1015	2	E08	APPLICATION SI
3	CDISCPILOT01	AE	01-701-1015	3	E06	DIARRHOEA
4	CDISCPILOT01	AE	01-701-1023	3	E10	ATRIOVENTRICU
5	CDISCPILOT01	AE	01-701-1023	1	E08	ERYTHEMA
6	CDISCPILOT01	AE	01-701-1023	2	E09	ERYTHEMA

[4]:
```
summary(ae)
```

```
  STUDYID             DOMAIN             USUBJID              AESEQ      
Length:1191        Length:1191        Length:1191        Min.   : 1.00  
Class :character   Class :character   Class :character   1st Qu.: 2.00  
Mode  :character   Mode  :character   Mode  :character   Median : 4.00  
                                                         Mean   : 4.53  
                                                         3rd Qu.: 6.00  
                                                         Max.   :23.00  

   AESPID             AETERM             AELLT              AELLTCD      
Length:1191        Length:1191        Length:1191        Mode:logical  
Class :character   Class :character   Class :character   NA's:1191     
Mode  :character   Mode  :character   Mode  :character                 

  AEDECOD             AEPTCD             AEHLT              AEHLTCD      
Length:1191        Mode:logical       Length:1191        Mode:logical  
Class :character   NA's:1191          Class :character   NA's:1191     
Mode  :character                      Mode  :character                 

   AEHLGT             AEHLGTCD           AEBODSYS           AEBDSYCD      
Length:1191        Mode:logical       Length:1191        Mode:logical  
Class :character   NA's:1191          Class :character   NA's:1191     
Mode  :character                      Mode  :character                 

   AESOC              AESOCCD            AESEV              AESER        
Length:1191        Mode:logical       Length:1191        Length:1191   
Class :character   NA's:1191          Class :character   Class :character
Mode  :character                      Mode  :character   Mode  :character
```

```
 AEACN            AEREL              AEOUT              AESCAN
Mode:logical   Length:1191        Length:1191        Length:1191
NA's:1191      Class :character   Class :character   Class :character
               Mode  :character   Mode  :character   Mode  :character
```

```
  AESCONG            AESDISAB           AESDTH             AESHOSP
Length:1191        Length:1191        Length:1191        Length:1191
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
  AESLIFE            AESOD              AEDTC              AESTDTC
Length:1191        Length:1191        Length:1191        Length:1191
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
  AEENDTC              AESTDY            AEENDY
Length:1191        Min.   :-277.00   Min.   : -2.00
Class :character   1st Qu.:  15.00   1st Qu.: 27.00
Mode  :character   Median :  32.00   Median : 53.00
                   Mean   :  45.83   Mean   : 67.14
                   3rd Qu.:  63.00   3rd Qu.:101.25
                   Max.   : 366.00   Max.   :211.00
                   NA's   :26        NA's   :473
```

[5]:
```
head(tibble::as_tibble(ae))
```

A tibble: 6 × 35

STUDYID <chr>	DOMAIN <chr>	USUBJID <chr>	AESEQ <int>	AESPID <chr>	AETERM <chr>
CDISCPILOT01	AE	01-701-1015	1	E07	APPLICATION SITE ERY
CDISCPILOT01	AE	01-701-1015	2	E08	APPLICATION SITE PRU
CDISCPILOT01	AE	01-701-1015	3	E06	DIARRHOEA
CDISCPILOT01	AE	01-701-1023	3	E10	ATRIOVENTRICULAR BL
CDISCPILOT01	AE	01-701-1023	1	E08	ERYTHEMA
CDISCPILOT01	AE	01-701-1023	2	E09	ERYTHEMA

[6]:
```
dplyr::glimpse(ae)
```

```
Rows: 1,191
Columns: 35
$ STUDYID  <chr> "CDISCPILOT01", "CDISCPILOT01",
"CDISCPILOT01", "CDISCPILOT01…
$ DOMAIN   <chr> "AE", "AE", "AE", "AE", "AE", "AE", "AE",
"AE", "AE", "AE", "…
$ USUBJID  <chr> "01-701-1015", "01-701-1015", "01-701-1015",
"01-701-1023", "…
$ AESEQ    <int> 1, 2, 3, 3, 1, 2, 4, 1, 2, 1, 2, 4, 1, 2, 3,
4, 10, 3, 1, 9, …
$ AESPID   <chr> "E07", "E08", "E06", "E10", "E08", "E09",
"E08", "E04", "E05"…
$ AETERM   <chr> "APPLICATION SITE ERYTHEMA", "APPLICATION
SITE PRURITUS", "DI…
$ AELLT    <chr> "APPLICATION SITE REDNESS", "APPLICATION
SITE ITCHING", "DIAR…
$ AELLTCD  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ AEDECOD  <chr> "APPLICATION SITE ERYTHEMA", "APPLICATION
SITE PRURITUS", "DI…
$ AEPTCD   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ AEHLT    <chr> "HLT_0617", "HLT_0317", "HLT_0148",
"HLT_0415", "HLT_0284", "…
$ AEHLTCD  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ AEHLGT   <chr> "HLGT_0152", "HLGT_0338", "HLGT_0588",
"HLGT_0086", "HLGT_019…
$ AEHLGTCD <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ AEBODSYS <chr> "GENERAL DISORDERS AND ADMINISTRATION SITE
CONDITIONS", "GENE…
$ AEBDSYCD <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ AESOC    <chr> "GENERAL DISORDERS AND ADMINISTRATION SITE
CONDITIONS", "GENE…
$ AESOCCD  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ AESEV    <chr> "MILD", "MILD", "MILD", "MILD", "MILD",
"MODERATE", "MILD", "…
$ AESER    <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", "N", "…
$ AEACN    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, N…
$ AEREL    <chr> "PROBABLE", "PROBABLE", "REMOTE",
"POSSIBLE", "POSSIBLE", "PR…
$ AEOUT    <chr> "NOT RECOVERED/NOT RESOLVED", "NOT
RECOVERED/NOT RESOLVED", "…
```

```
$ AESCAN   <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", "…
$ AESCONG  <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", "…
$ AESDISAB <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", "…
$ AESDTH   <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", "…
$ AESHOSP  <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", "…
$ AESLIFE  <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", "…
$ AESOD    <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N",
"N", "N", "N", "…
$ AEDTC    <chr> "2014-01-16", "2014-01-16", "2014-01-16",
"2012-08-27", "2012…
$ AESTDTC  <chr> "2014-01-03", "2014-01-03", "2014-01-09",
"2012-08-26", "2012…
$ AEENDTC  <chr> NA, NA, "2014-01-11", NA, "2012-08-30", NA,
"2012-08-30", NA,…
$ AESTDY   <int> 2, 2, 8, 22, 3, 3, 3, 3, 21, 58, 125, 27, 1,
1, 23, 52, 52, 5…
$ AEENDY   <int> NA, NA, 10, NA, 26, NA, 26, NA, NA, NA, NA,
NA, 1, 1, NA, NA,…
```

1.2 DROP und KEEP: Variablenselektion

Das Paket dplyr hat den Befehl **select()**, mit dem Variablen eingeschlossen und ausgeschlossen werden können.

[7]:
```
# select() wählt Variablen aus
# %>% wird als Piping-Operator bezeichnet, das Ergebnis links wird als Eingabe␣
 ↪rechts an erster Stelle eingegeben.
# select() in dieser Form entspricht dem KEEP.
ae1 <- ae %>% select(DOMAIN, USUBJID, AEDECOD, AEBODSYS)
head(ae1)
```

A data.frame: 6 × 4

	DOMAIN <chr>	USUBJID <chr>	AEDECOD <chr>	AE <ch
1	AE	01-701-1015	APPLICATION SITE ERYTHEMA	GE
2	AE	01-701-1015	APPLICATION SITE PRURITUS	GE
3	AE	01-701-1015	DIARRHOEA	GA
4	AE	01-701-1023	ATRIOVENTRICULAR BLOCK SECOND DEGREE	CA
5	AE	01-701-1023	ERYTHEMA	SK
6	AE	01-701-1023	ERYTHEMA	SK

[8]:
```
# Die Domain soll ebenfalls entfernt werden. Mehrere Variablen werden als␣
 ↪Vektor übergeben -c(DOMAIN, AEDECOD)
```

```
# select mit Minus-Zeichen entspricht dem DROP.
ae2 <- ae1 %>% select(-DOMAIN)
head(ae2)
```

A data.frame: 6 × 3

	USUBJID <chr>	AEDECOD <chr>	AEBODSYS <chr>
1	01-701-1015	APPLICATION SITE ERYTHEMA	GENERAL DISO
2	01-701-1015	APPLICATION SITE PRURITUS	GENERAL DISO
3	01-701-1015	DIARRHOEA	GASTROINTES
4	01-701-1023	ATRIOVENTRICULAR BLOCK SECOND DEGREE	CARDIAC DISO
5	01-701-1023	ERYTHEMA	SKIN AND SUB
6	01-701-1023	ERYTHEMA	SKIN AND SUB

[9]:
```
ae %>% select(ends_with("DY")) %>% head()
```

A data.frame: 6 × 2

	AESTDY <int>	AEENDY <int>
1	2	NA
2	2	NA
3	8	10
4	22	NA
5	3	26
6	3	NA

[10]:
```
ae %>% select(starts_with("AELLT")) %>% head()
```

A data.frame: 6 × 2

	AELLT <chr>	AELLTCD <lgl>
1	APPLICATION SITE REDNESS	NA
2	APPLICATION SITE ITCHING	NA
3	DIARRHEA	NA
4	AV BLOCK SECOND DEGREE	NA
5	ERYTHEMA	NA
6	LOCALIZED ERYTHEMA	NA

[11]:
```
ae %>% select(matches(".ST.")) %>% head()
```

A data.frame: 6 × 2

	AESTDTC <chr>	AESTDY <int>
1	2014-01-03	2
2	2014-01-03	2
3	2014-01-09	8
4	2012-08-26	22
5	2012-08-07	3
6	2012-08-07	3

1.3 WHERE: Selektion von Beobachtungen

Bedingungen können mit UND (&) und/oder ODER (|) verknüpft und mit Klammern priorisiert werden. Weitere boolsche Operatoren sind !, xor, any, all.

Folgende Operatoren stehen zur Verfügung:

- < : Kleiner als
- > : Größer als
- == : Gleich
- <= : Kleiner oder gleich
- >= : Größer oder gleich
- != : Ungleich
- %in% : ist enthalten in
- is.na : ist missing
- !is.na : ist nicht missing
- is.null : ist null
- !is.null : ist nicht null

[12]:
```
ae1 %>% filter(AEDECOD == "ERYTHEMA") %>% head()
```

A data.frame: 6 × 4

	DOMAIN <chr>	USUBJID <chr>	AEDECOD <chr>	AEBODSYS <chr>
1	AE	01-701-1023	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS
2	AE	01-701-1023	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS
3	AE	01-701-1023	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS
4	AE	01-701-1097	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS
5	AE	01-701-1111	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS
6	AE	01-701-1111	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS

[13]:
```
ae1 %>% filter(AEDECOD == "ERYTHEMA" & USUBJID == "01-701-1023") %>% head()
```

A data.frame: 3 × 4

	DOMAIN <chr>	USUBJID <chr>	AEDECOD <chr>	AEBODSYS <chr>
1	AE	01-701-1023	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS
2	AE	01-701-1023	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS
3	AE	01-701-1023	ERYTHEMA	SKIN AND SUBCUTANEOUS TISSUE DIS

[14]:
```
ae1 %>% filter(AEDECOD == "ERYTHEMA" | USUBJID == "01-701-1023") %>% head()
```

A data.frame: 6 × 4

	DOMAIN <chr>	USUBJID <chr>	AEDECOD <chr>	AE <ch
1	AE	01-701-1023	ATRIOVENTRICULAR BLOCK SECOND DEGREE	CA
2	AE	01-701-1023	ERYTHEMA	SKI
3	AE	01-701-1023	ERYTHEMA	SKI
4	AE	01-701-1023	ERYTHEMA	SKI
5	AE	01-701-1097	ERYTHEMA	SKI
6	AE	01-701-1111	ERYTHEMA	SKI

[15]:
```
# Bestimmte Beobachtungen herausschneiden
head(ae1)
ae2 <- ae1 %>% slice(2:4)
ae2
```

A data.frame: 6 × 4

	DOMAIN <chr>	USUBJID <chr>	AEDECOD <chr>	AE… <ch…
1	AE	01-701-1015	APPLICATION SITE ERYTHEMA	GEN…
2	AE	01-701-1015	APPLICATION SITE PRURITUS	GEN…
3	AE	01-701-1015	DIARRHOEA	GAS…
4	AE	01-701-1023	ATRIOVENTRICULAR BLOCK SECOND DEGREE	CAR…
5	AE	01-701-1023	ERYTHEMA	SKI…
6	AE	01-701-1023	ERYTHEMA	SKI…

A data.frame: 3 × 4

DOMAIN <chr>	USUBJID <chr>	AEDECOD <chr>	AEBOD… <chr>
AE	01-701-1015	APPLICATION SITE PRURITUS	GENER…
AE	01-701-1015	DIARRHOEA	GASTR…
AE	01-701-1023	ATRIOVENTRICULAR BLOCK SECOND DEGREE	CARDI…

1.3.1 MERGE BY: Verknüpfen von Datensätzen

Vier grundsätzliche Typen:

- left_join
- right_join
- inner_join
- full_join

TODO: Visualisierung über Venn-Diagramme ergänzen

[16]:
```
# Die erste 4 Studienteilnehmer im Datensatz suchen
head(unique(ae$USUBJID), 4)
```

1. '01-701-1015' 2. '01-701-1023' 3. '01-701-1028' 4. '01-701-1034'

[17]:
```
# Diesen willkürlich zu Demo-Zwecken eine Therapie zuweisen und einen weiteren␣
 ↪Patienten ergänzen, der im Datensatz
# ae nicht vorkommt.
# rep() ist eine Funktion zur Wiederholung eines Vektors.
# seq() gehört zu dieser Funktion dazu.
USUBJID = c('01-701-1015', '01-701-1023', '01-701-1028', '01-701-1034',␣
 ↪'01-702-1001')
TRTP = c(rep(c("TRT_A", "TRT_B"), 2), "TRT_C")

Applied_TRT <- data.frame(USUBJID, TRTP)
Applied_TRT
```

A data.frame: 5 × 2

USUBJID	TRTP
<chr>	<chr>
01-701-1015	TRT_A
01-701-1023	TRT_B
01-701-1028	TRT_A
01-701-1034	TRT_B
01-702-1001	TRT_C

[18]:
```
# Alle Patienten, die ein Treatment haben.
inner_join(ae, Applied_TRT, by = c("USUBJID")) %>% select(USUBJID, AETERM,␣
↪TRTP) # %>% select(USUBJID) %>% unique() %>% count()
```

A data.frame: 11 × 3

USUBJID	AETERM	TRTP
<chr>	<chr>	<chr>
01-701-1015	APPLICATION SITE ERYTHEMA	TRT_A
01-701-1015	APPLICATION SITE PRURITUS	TRT_A
01-701-1015	DIARRHOEA	TRT_A
01-701-1023	ATRIOVENTRICULAR BLOCK SECOND DEGREE	TRT_B
01-701-1023	ERYTHEMA	TRT_B
01-701-1023	ERYTHEMA	TRT_B
01-701-1023	ERYTHEMA	TRT_B
01-701-1028	APPLICATION SITE ERYTHEMA	TRT_A
01-701-1028	APPLICATION SITE PRURITUS	TRT_A
01-701-1034	APPLICATION SITE PRURITUS	TRT_B
01-701-1034	FATIGUE	TRT_B

[19]:
```
# Alle Patienten, die kein Treatment haben.
# Strings haben einen NA Wert.
left_join(ae, Applied_TRT, by = c("USUBJID")) %>% select(USUBJID, AETERM, TRTP)␣
↪%>% filter(is.na(TRTP)) %>% select(USUBJID) %>% unique() %>% count()
```

A data.frame: 1 × 1

n
<int>
221

[20]:
```
# Patienten mit Treatment und ohne AE.
right_join(ae, Applied_TRT, by = c("USUBJID")) %>% select(USUBJID, AETERM,␣
↪TRTP) %>% filter(is.na(AETERM))
```

A data.frame: 1 × 3

USUBJID	AETERM	TRTP
<chr>	<chr>	<chr>
01-702-1001	NA	TRT_C

[ ]:
```

```