R Kurs Unterlagen

Anna-Lena Schubert, Jan Goettmann, Jose Carlos Garcia Alanis, Meike Steinhilber, Cordula Hu

2021-10-13

Inhaltsverzeichnis

Kapitel 1

Über dieses Buch

TEXT

Kapitel 2

Einführung

(Anna-Lena)

Kapitel 3

Datenstruktur

(Florian)

3.1 Einführung in Dplyr und tidyverse

Dplyr ist Teil des tidyverse Packages und ermöglicht es, Daten sehr einfach zu manipulieren und in eine Form zu bringen, um diese dann zu analysieren. Um das zu tun greifen wir auf den Star Wars Datensatz zurück, den das dplyr Package mitliefert:

```
# Lest die Daten bitte ein, der Datensatz heisst "starwars.RDS" und befindet sich in eurem Projek

starwars <- readRDS("starwars.RDS")
```

Der Datensatz enthält Informationen über unsere Star Wars Helden, ähnlich dem Datensatz, den wir uns in der letzten Sitzung ausgedacht haben:

```
head(starwars,5) # Wir lassen uns erstmal die ersten 5 Zeilen des Datensatzes ausgeben
```

```
## # A tibble: 5 x 11
##   name           height  mass hair_color skin_color  eye_color   Age sex   gender
##   <chr>           <int> <dbl> <fct>      <fct>       <fct>     <dbl> <fct> <fct>
## 1 Luke Skywalker    172    77 blond      fair        blue       19   male  mascu~
## 2 C-3PO             167    75 <NA>       gold        yellow    112   none  mascu~
## 3 R2-D2              96    32 <NA>       white, blue red        33   none  mascu~
## 4 Darth Vader       202   136 none       white       yellow     41.9 male  mascu~
## 5 Leia Organa       150    49 brown      light       brown      19   fema~ femin~
## # ... with 2 more variables: homeworld <chr>, species <chr>
```

Bevor wir einsteigen, schaut euch an, wie die einzelnen Variablen im Datensatz verteilt sind. Benutzt dazu den den `summary()` Befehl, was fällt euch auf ?

```
summary(starwars)
```

```
##      name               height           mass           hair_color   skin_color
##  Length:87          Min.   : 66.0   Min.   :  15.00   none   :37   fair   :17
##  Class :character   1st Qu.:167.0   1st Qu.:  55.60   brown  :18   light  :11
##  Mode  :character   Median :180.0   Median :  79.00   black  :13   dark   : 6
##                     Mean   :174.4   Mean   :  97.31   white  : 4   green  : 6
##                     3rd Qu.:191.0   3rd Qu.:  84.50   blond  : 3   grey   : 6
##                     Max.   :264.0   Max.   :1358.00   (Other): 7   pale   : 5
##                     NA's   :6       NA's   :28        NA's   : 5   (Other):36
##    eye_color       Age                       sex           gender
##  brown  :21   Min.   :  8.00   female        :16   feminine :17
##  blue   :19   1st Qu.: 35.00   hermaphroditic: 1   masculine:66
##  yellow :11   Median : 52.00   male          :60   NA's     : 4
##  black  :10   Mean   : 87.57   none          : 6
##  orange : 8   3rd Qu.: 72.00   NA's          : 4
##  red    : 5   Max.   :896.00
##  (Other):13   NA's   :44
##   homeworld          species
##  Length:87          Length:87
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

3.2 Dplyr: Die wichtigsten Befehle

- Filtern von Beobachtungen nach Wert (`filter()`).
- Reihen neu Sortieren (`arrange()`).
- Auswahl von Variablen nach Name (`select()`).
- Erstellen von neuen Variablen aus bereits existierenden (`mutate()`).
- Viele Werte zu einem einzelnen Wert zusammenfassen (`summarise()`).

Der vielleicht wichtigste Befehl ist der `group_by()` Befehl, mit dem Ihr die oben genannten Befehle auf einzelne Gruppen innerhalb eines Datensatzes anwenden könnt.

Diese 6 sogennaten “Verben” bilden die Grundlage für tidyverse.Damit ist es möglichmehrere einfache Schritte miteinander zu verketten, um ein komplexes Ergebnis zu erzielen. Alles Befehle funktionieren auf die gleiche Art und Weise: