04 Datenmanipulation mit dplyR

Martin Hanewald

2019-02-28

Packages

```
library(tidyverse)
library(nycflights13)
library(rvest)
library(knitr)
library(DT)
```

Überblick

Basis Operationen von `dplyr`:

- `filter()` zum Filtern nach Werten
- `arrange()` zum Sortieren
- `select()` und `rename()` zum Auswählen von Spalten und Umbenennen
- `mutate()` und `transmute()` zum Erzeugen neuer Spalten
- `group_by()` zum Definieren einer Gruppierungsebene
- `summarise()`zum Aggregieren von Kennzahlen auf einer Gruppierungsebene

Demonstration mit Star Wars Charakteren:

Welche Charaktere spielen am häufigsten mit?

```
data(starwars)

starwars %>% select(name, films) %>%
    unnest() %>%
    group_by(name) %>%
    count() %>% arrange(desc(n)) %>%
    head(10) %>%
    kable()
```

name	n
R2-D2	7
C-3PO	6
Obi-Wan Kenobi	6
Chewbacca	5
Leia Organa	5
Luke Skywalker	5
Palpatine	5
Yoda	5

name	n
Darth Vader	4
Han Solo	4

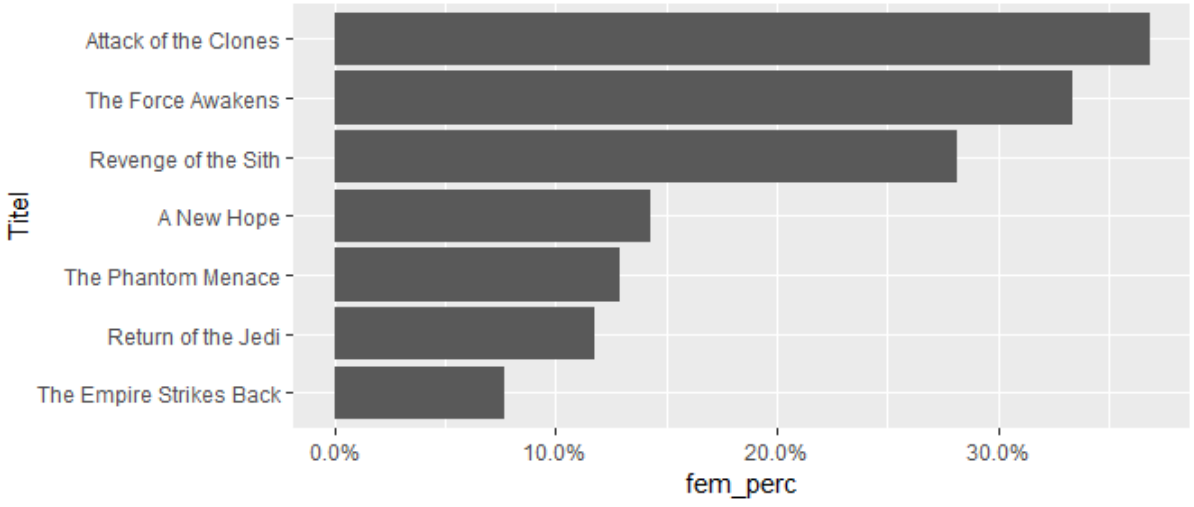
Berechne die Frauenquote pro Film ?

```
ans <- starwars %>% select(name, gender, films) %>% unnest %>%
    group_by(films, gender) %>% count() %>%
    filter(gender %in% c('female', 'male')) %>%
    spread(gender, n) %>%
    mutate(fem_perc = female / (female + male)) %>%
    select(films, fem_perc) %>%
    arrange(desc(fem_perc))

ans %>% kable()
```

films	fem_perc
Attack of the Clones	0.3684211
The Force Awakens	0.3333333
Revenge of the Sith	0.2812500
A New Hope	0.1428571
The Phantom Menace	0.1290323
Return of the Jedi	0.1176471
The Empire Strikes Back	0.0769231

```
ans %>%
    ggplot(aes(films %>% fct_reorder(fem_perc), fem_perc)) +
        geom_col() + coord_flip() + scale_y_continuous(labels=scales::percent)+
        labs(x='Titel', 'Anteil Frauen')
```



Aufgaben: Analysiere Flugdaten

Datensatz `flights` aus Package `nycflights13`.

```
data(flights)
```


Show 10 entries Search:

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time
1	2013	8	31	1100	1040	20	1214
2	2013	8	9	2116	1936	100	20
3	2013	5	31	954	1000	-6	1053
4	2013	12	19	717	715	2	908
5	2013	4	16	1803	1455	188	2101
6	2013	10	16	821	825	-4	1030
7	2013	5	15	629	630	-1	828
8	2013	7	19	1746	1655	51	2038
9	2013	11	27	1650	1650	0	1951
10	2013	1	22	1438	1429	9	1634

Showing 1 to 10 of 10 entries Previous 1 Next

Welches sind die beliebtesten Reiseziele?

Liste die Top 10

```
flights %>%
    group_by(dest) %>%
    count() %>%
    arrange(desc(n)) %>%
    head(10) %>%
    kable()
```

dest	n
ORD	17283
ATL	17215
LAX	16174
BOS	15508
MCO	14082
CLT	14064

dest	n
SFO	13331
FLL	12055
MIA	11728
DCA	9705

Welches sind die unpünktlichsten Fluggesellschaften

- Berechne die mittlere Verspätung im Verhältnis zur Strecke
- Sortiere nach schlechtesten Fluggesellschaften

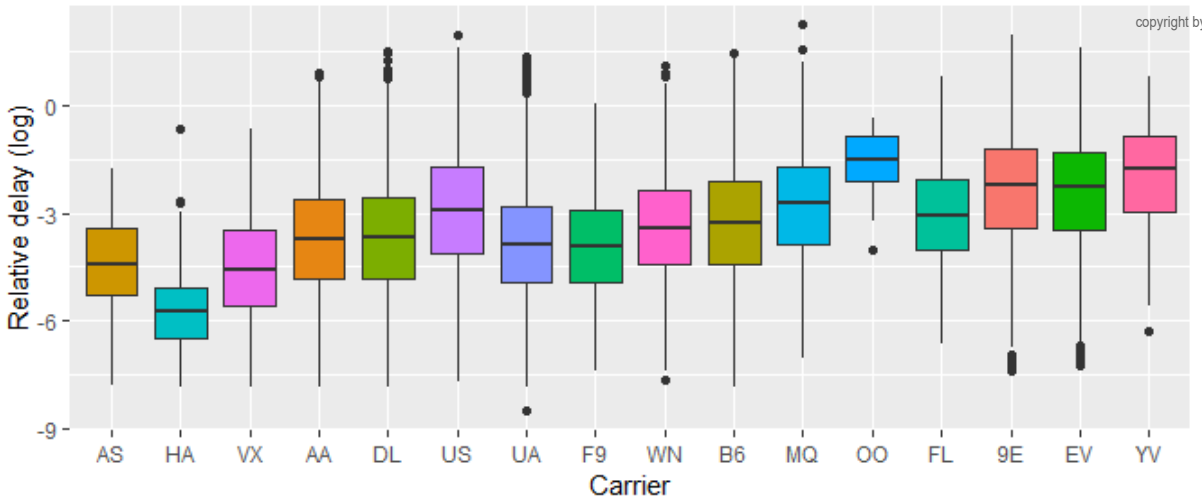
```
ans <- flights %>%
    mutate(tot_delay = dep_delay + arr_delay) %>%
    mutate(rel_delay = tot_delay / distance) %>%
    group_by(carrier) %>%
    summarise(rel_delay = mean(rel_delay, na.rm =T),
              count = n()) %>%
    arrange(desc(rel_delay))


ans %>% head(10) %>% kable()
```

carrier	rel_delay	count
YV	0.1125663	601
EV	0.0854153	54173
9E	0.0697272	18460
FL	0.0637883	3260
OO	0.0498792	32
MQ	0.0442328	26397
B6	0.0383656	54635
WN	0.0317897	12275
F9	0.0260012	685
UA	0.0172654	58665

```
# Darstellung als Boxplot
flights %>%
    mutate(tot_delay = dep_delay + arr_delay) %>%
    mutate(rel_delay = tot_delay / distance) %>%
    ggplot(aes(carrier %>% fct_reorder(rel_delay, mean, na.rm=T), log(rel_delay), fill=carrier)) +
    geom_boxplot() + labs(y='Relative delay (log)', x='Carrier') +
    theme(legend.position="none")
#> Warning in log(rel_delay): NaNs wurden erzeugt

#> Warning in log(rel_delay): NaNs wurden erzeugt
#> Warning: Removed 201717 rows containing non-finite values (stat_boxplot).
```



Bonus: Auflösung des IATA Codes in Namen

```
url <- "https://aspmhelp.faa.gov/index.php/ASQP_:_Carrier_Codes_And_Names"
carrier_codes <- url %>%
    read_html() %>%
    html_table() %>%
    .[[1]]

ans %>%
    left_join(carrier_codes, by=c('carrier'='IATA Carrier Code')) %>%
    kable()
```

carrier	rel_delay	count	ICAO Carrier Code	Carrier Name
YV	0.1125663	601	ASH	Mesa Airlines
EV	0.0854153	54173	CAA	Atlantic Southeast Airlines
9E	0.0697272	18460	FLG	Pinnacle Airlines
FL	0.0637883	3260	TRS	AirTran Airways
OO	0.0498792	32	SKW	SkyWest Airlines
MQ	0.0442328	26397	EGF	American Eagle
B6	0.0383656	54635	JBU	JetBlue Airways
WN	0.0317897	12275	SWA	Southwest Airlines
F9	0.0260012	685	FFT	Frontier
UA	0.0172654	58665	UAL	United Airlines
US	0.0168932	20536	USA	US Airways
DL	0.0132526	48110	DAL	Delta Air Lines
AA	0.0087874	32729	AAL	American Airlines
VX	0.0057073	5162	NA	NA
HA	-0.0004043	342	HAL	Hawaiian Airlines
AS	-0.0017070	714	ASA	Alaska Airlines