The Oscars Award

Explorative Datenanalyse & Visualisierung

Mathias Petak, Ivo Otero

4 1 2022

Contents

Aufgabenstellung

Datenaufbereitung und Explorative Datenanalyse, speziell Visualisierung

Packages laden

```
library(tidyverse)
library(ggplot2)
library(ggmosaic)
library(pander)
```

Daten einlesen

```
oscars_data <- read.csv(file = '../data/the_oscar_award.csv', header = TRUE, sep = ",", encoding = "UTF
```

Daten aufbereiten

```
oscars_tbl <- as_tibble(oscars_data)
oscars_tbl$winner <- as.factor(oscars_tbl$winner)
oscars_tbl$category <- as.factor(oscars_tbl$category)
head(oscars_tbl)
```

```
## # A tibble: 6 x 7
##   year_film year_ceremony ceremony category      name                film  winner
##       <int>         <int>    <int> <fct>         <chr>               <chr> <fct>
## 1      1927          1928        1 ACTOR         Richard Barthelmess The ~ False
## 2      1927          1928        1 ACTOR         Emil Jannings       The ~ True
## 3      1927          1928        1 ACTRESS       Louise Dresser      A Sh~ False
## 4      1927          1928        1 ACTRESS       Janet Gaynor        7th ~ True
## 5      1927          1928        1 ACTRESS       Gloria Swanson      Sadi~ False
## 6      1927          1928        1 ART DIRECTION Rochus Gliese       Sunr~ False
```

Daten untersuchen

```
summary(oscars_tbl)
```

```
##    year_film    year_ceremony     ceremony
##  Min.   :1927   Min.   :1928   Min.   : 1.0
##  1st Qu.:1951   1st Qu.:1952   1st Qu.:24.0
##  Median :1974   Median :1975   Median :47.0
##  Mean   :1974   Mean   :1975   Mean   :47.5
##  3rd Qu.:1998   3rd Qu.:1999   3rd Qu.:71.0
##  Max.   :2019   Max.   :2020   Max.   :92.0
##
##                          category        name               film
##  DIRECTING                   : 449   Length:10395       Length:10395
##  FILM EDITING                : 430   Class :character   Class :character
##  ACTOR IN A SUPPORTING ROLE  : 420   Mode  :character   Mode  :character
##  ACTRESS IN A SUPPORTING ROLE: 420
##  DOCUMENTARY (Short Subject) : 368
##  DOCUMENTARY (Feature)       : 335
##  (Other)                     :7973
##    winner
##  False:8038
##  True :2357
##
##
```

```
##
##
##
```

Der Datensatz besteht aus 7 Variablen (3 metrischen und 4 kategorialen) und 10395 Beobachtungen. Eine Beobachtung beschreibt jeweils eine Nominierung und das ihr zugehörige Erscheinungsjahr des Films (year_film), das Jahr der Zeremonie (year_ceremony), die nummerische Reihenfolge der Zeremonie (ceremony), die Kategorie des Preises (category), den Namen des Nominierten (name), den betreffenden Film (film) und ob die Nominierung letztendlich gewonnen hat (winner).

Fehlende Werte

```
sum(is.na(oscars_tbl))
```

```
## [1] 304
```

```
oscars_tbl %>%
  filter(is.na(film))
```

```
## # A tibble: 304 x 7
##    year_film year_ceremony ceremony category                name    film  winner
##        <int>         <int>    <int> <fct>                   <chr>   <chr> <fct>
##  1      1927          1928        1 ENGINEERING EFFECTS     "Ralph~ <NA>  False
##  2      1927          1928        1 ENGINEERING EFFECTS     "Nugen~ <NA>  False
##  3      1927          1928        1 WRITING (Title Writing) "Josep~ <NA>  True
##  4      1927          1928        1 WRITING (Title Writing) "Georg~ <NA>  False
##  5      1927          1928        1 SPECIAL AWARD           " Warn~ <NA>  True
##  6      1927          1928        1 SPECIAL AWARD           " Char~ <NA>  True
##  7      1930          1931        4 SOUND RECORDING         "Samue~ <NA>  False
##  8      1930          1931        4 SOUND RECORDING         "Metro~ <NA>  False
##  9      1930          1931        4 SOUND RECORDING         "Param~ <NA>  True
## 10      1930          1931        4 SOUND RECORDING         "RKO R~ <NA>  False
## # ... with 294 more rows
```

Es gibt 304 fehlende Werte, alle betreffen die Variable “name”.

Häufigkeit der Nominierungen nach Jahren (year_film, year_ceremony)

Es wird untersucht, ob year_ceremony nur auf Filme aus dem Vorjahr zurückgreift (year_film):

```
matching_yrs <- 0

for (row in 1:nrow(oscars_tbl)){
  if (oscars_tbl$year_film[row]+1 == oscars_tbl$year_ceremony[row])
    matching_yrs <- matching_yrs+1
}

matching_yrs # Filme aus dem Vorjahr
```

```
## [1] 10395
```

```
matching_yrs/nrow(oscars_tbl) # Filme aus dem Vorjahr / Alle Beobachtungen
```

```
## [1] 1
```

Für alle 10395 Beobachtungen wurden Filme aus dem Vorjahr nominiert.

Anzahl der Nominierungen

```
cnt_yrs <- count(oscars_tbl, year_ceremony) %>% arrange(desc(n))
c_yrs_mean <- mean(cnt_yrs$n)
cnt_yrs
```

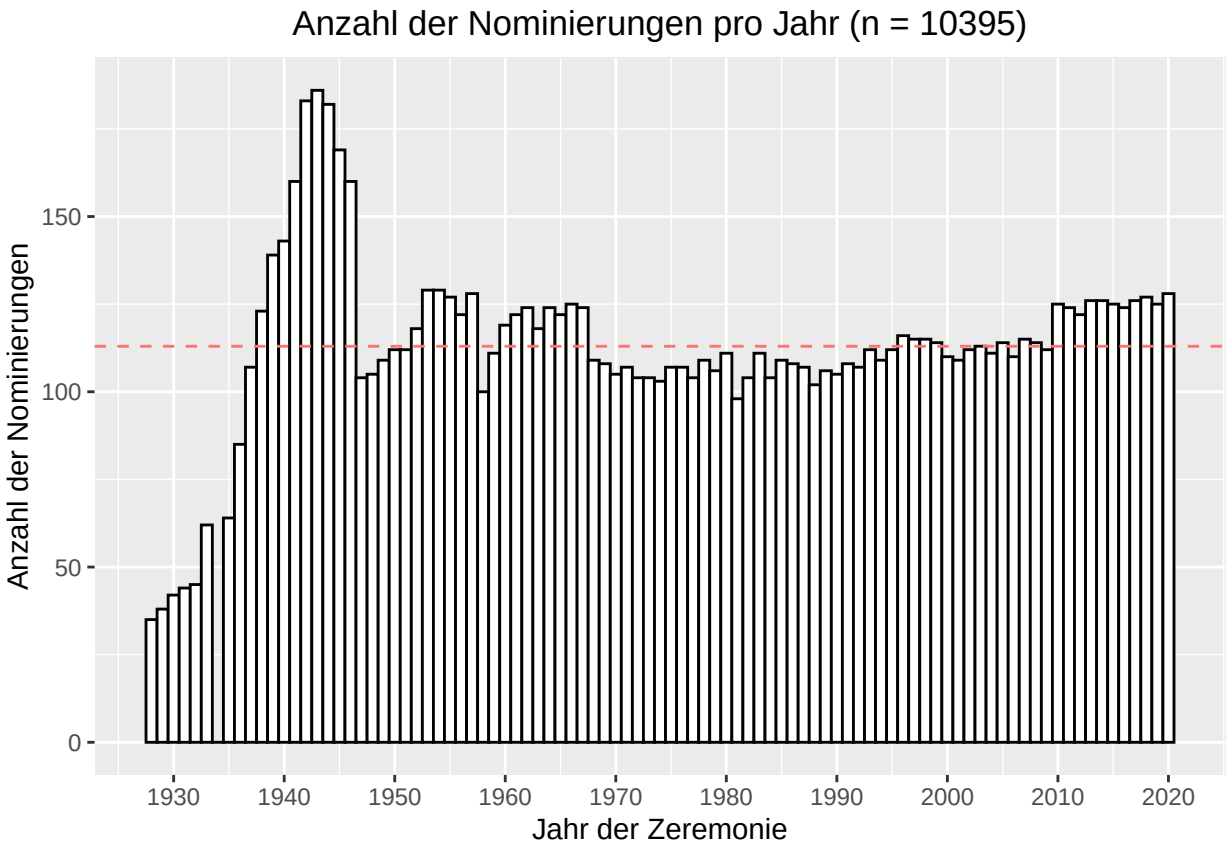
```
## # A tibble: 92 x 2
##    year_ceremony     n
##            <int> <int>
##  1          1943   186
##  2          1942   183
##  3          1944   182
##  4          1945   169
##  5          1941   160
##  6          1946   160
##  7          1940   143
##  8          1939   139
##  9          1953   129
## 10          1954   129
## # ... with 82 more rows
```

```
c_yrs_mean
```

```
## [1] 112.9891
```

Der Mittelwert der Nominierungen liegt in etwa bei 113.

```
ggplot(oscars_tbl, aes(x = year_ceremony)) +
  geom_histogram(color="black", fill="white", binwidth = 1) +
  scale_x_continuous(breaks=c(1930, 1940, 1950, 1960, 1970, 1980, 1990, 2000, 2010, 2020)) +
  labs(title="Anzahl der Nominierungen pro Jahr (n = 10395)",x="Jahr der Zeremonie", y =
         "Anzahl der Nominierungen") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_hline(aes(yintercept=c_yrs_mean, color="red"),
             linetype="dashed", show.legend = FALSE)
```



Es ist auffällig, dass es im Jahr 1934 keine Nominierung gab und daher vermutlich in diesem Jahr keine Oscar-Verleihung stattgefunden hat. Die Anzahl der Nominierungen erreichte im Jahr 1943 ihren Höchstwert (186 Nominierungen) und pendelte sich seit 1946 im Bereich zwischen 100 und 130 Nominierungen pro Jahr ein.

Kategorien

```
biggest_categories <- oscars_tbl %>%
  mutate(category = fct_lump(category, n = 10)) %>%
  count(category, sort = TRUE)

biggest_categories
```

```
## # A tibble: 11 x 2
##    category                          n
##    <fct>                         <int>
##  1 Other                          6700
##  2 DIRECTING                       449
##  3 FILM EDITING                    430
##  4 ACTOR IN A SUPPORTING ROLE      420
##  5 ACTRESS IN A SUPPORTING ROLE    420
##  6 DOCUMENTARY (Short Subject)     368
##  7 DOCUMENTARY (Feature)           335
##  8 BEST PICTURE                    333
```

```
##  9 CINEMATOGRAPHY                318
## 10 FOREIGN LANGUAGE FILM         315
## 11 ART DIRECTION                 307
```

```
sum(biggest_categories$n[2:11])
```

```
## [1] 3695
```

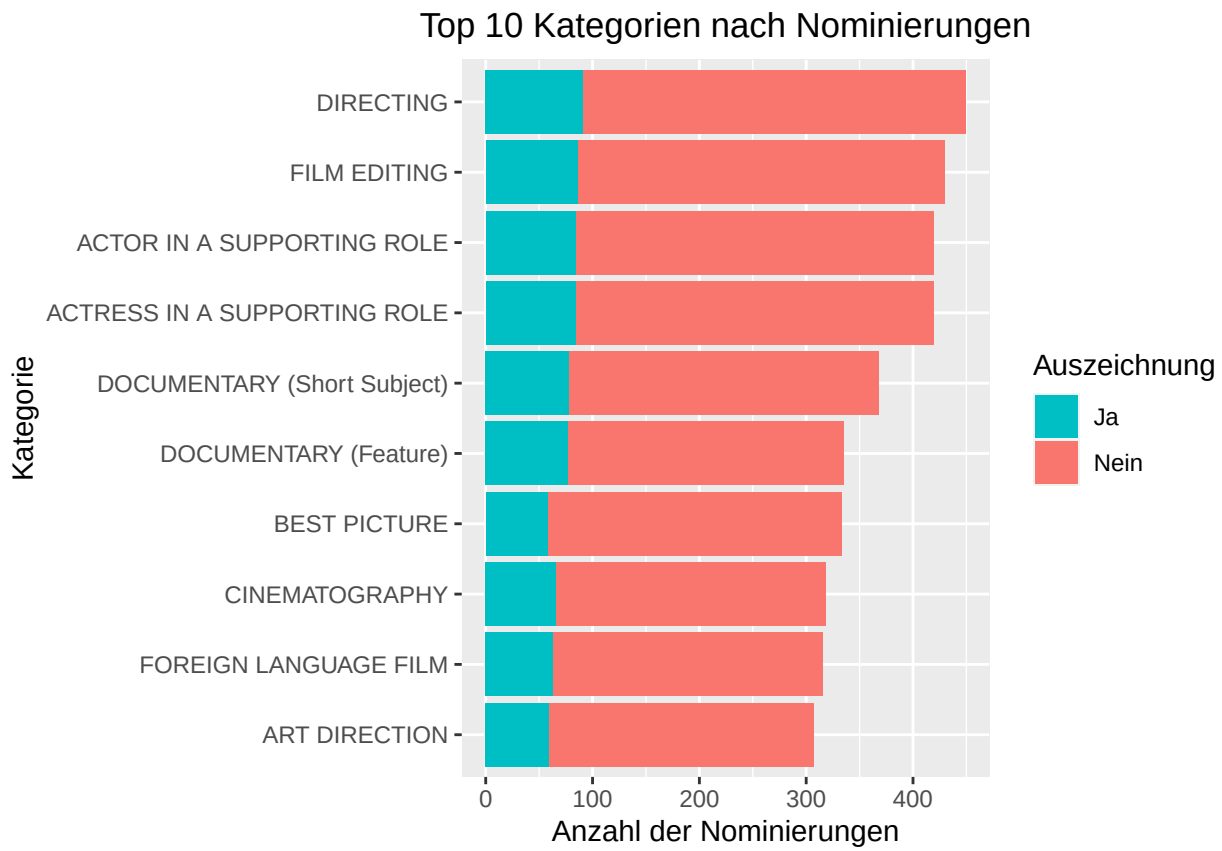
6700 Nominierungen entfallen auf alle anderen Kategorien zusammengerechnet. 3695 Nominierungen entfallen auf die Top 10.

```
sum(biggest_categories$n[2:11])/nrow(oscars_tbl) # Prozentuelle Anzahl der Nomierungen von Top 10 Kateg
```

```
## [1] 0.3554594
```

Die Top 10 Kategorien machen in etwa 35,5% aller Nominierungen aus.

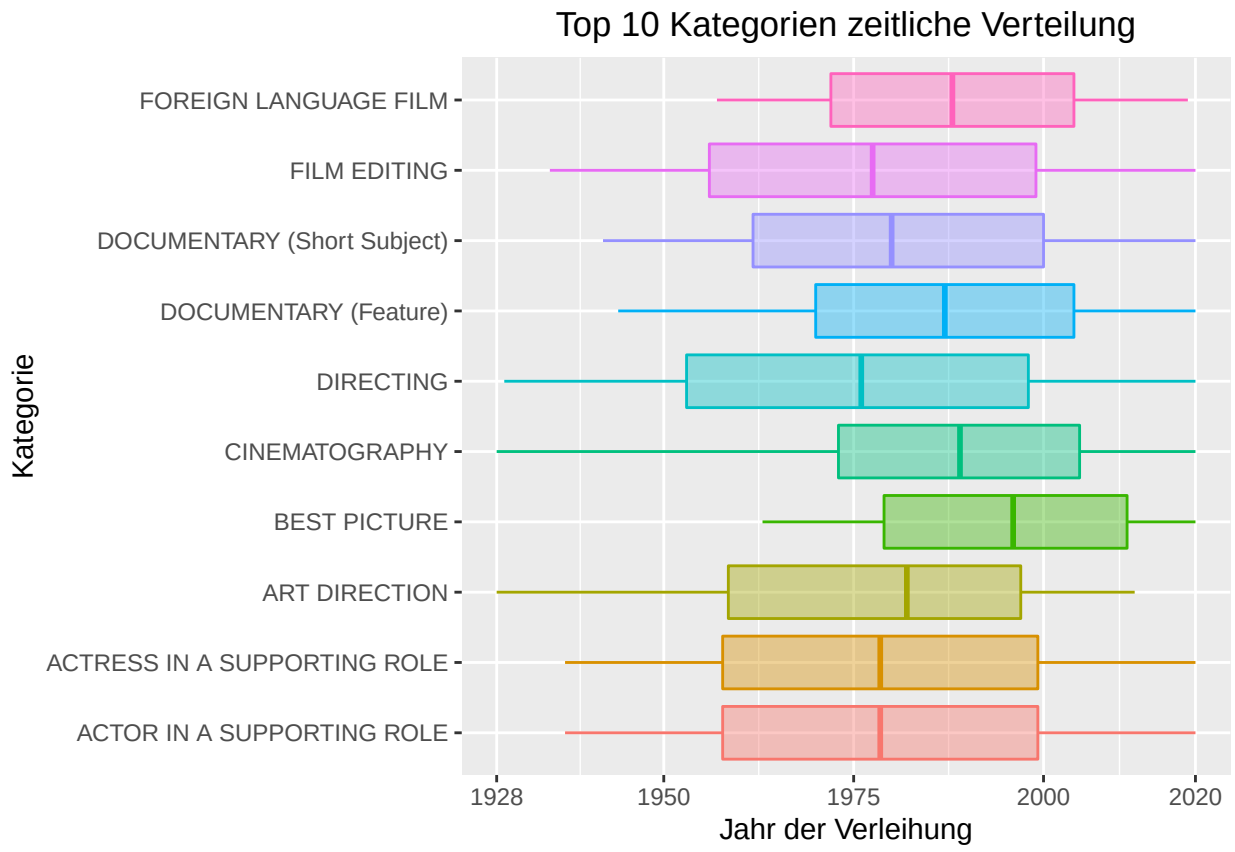
```
top_categories_filtered <- filter(oscars_tbl, category %in% biggest_categories$category)

ggplot(top_categories_filtered, aes(y = fct_rev(fct_infreq(category)), fill = winner)) +
  geom_bar() + labs(title = "Top 10 Kategorien nach Nominierungen", x = "Anzahl der Nominierungen",
                    y = "Kategorie") +
  scale_fill_discrete(guide = guide_legend(reverse=TRUE), name = "Auszeichnung", labels = c("Nein", "Ja
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_text(hjust = 0.5))
```



Die meisten Nominierungen (mit über 400) gibt es in den Kategorien “Directing”, “Film Editing”, sowie “Actor in a supporting Role”, “Actress in a supporting role”. Danach folgen “Documentary (Short Subject)”, “Documentary (Feature)”, “Best Picture”, “Cinematography”, “Foreign Language Film” und “Art Direction”.

Die Anzahl der Siege ist bei allen Kategorien weit unter der Hälfte der Nominierungen. Die Kategorie “Best Picture” hat von allen genannten die wenigsten Auszeichnungen, jedoch die 7. meisten Nominierungen.

```
ggplot(top_categories_filtered, aes(category, year_ceremony)) +
  geom_boxplot(aes(colour = category, fill = after_scale(alpha(colour, 0.4)))) +
  coord_flip() +
  scale_y_continuous(breaks = c(1928,1950, 1975, 2000, 2020)) +
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5)) + labs(title = "Top 10 Kategor
```



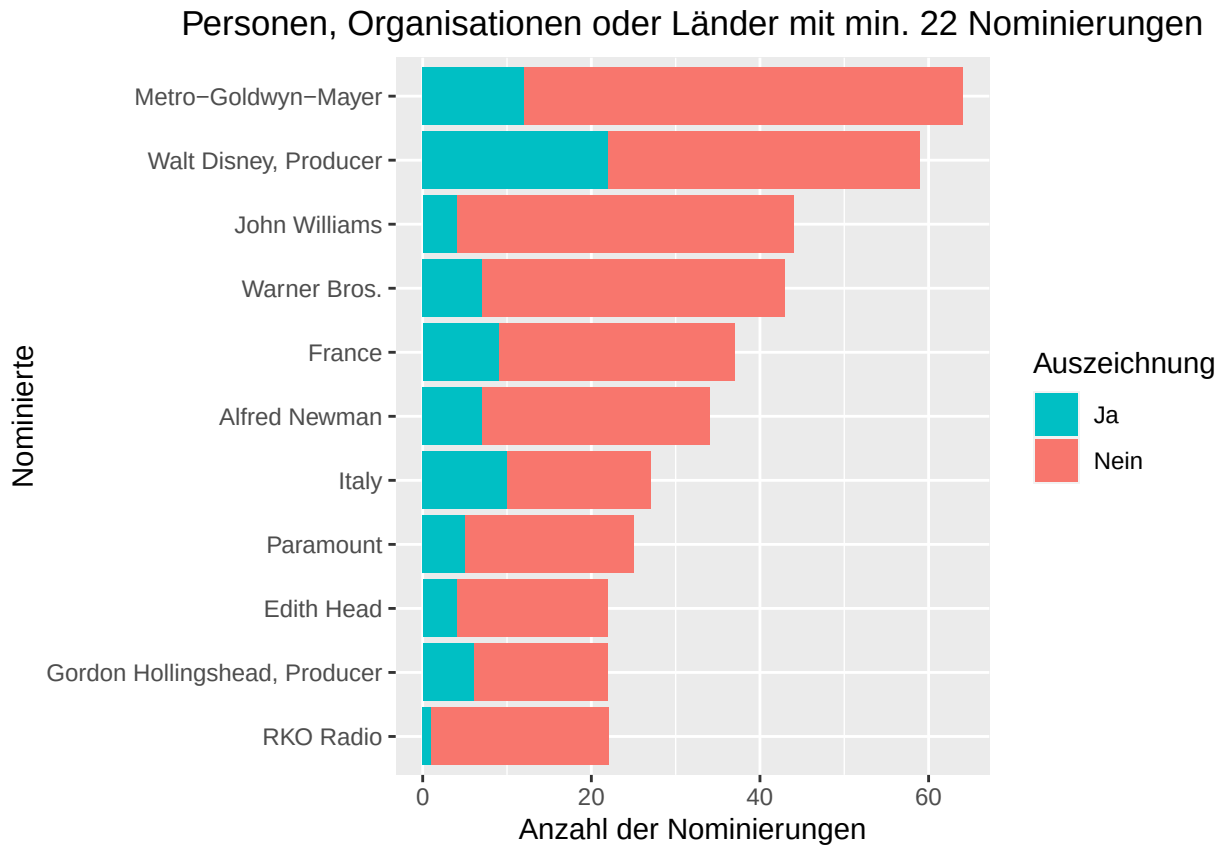
Cinematography ist die einzige Kategorie die seit Beginn der Oscar-Verleihungen bis zur Gegenwart (2020) immer vergeben wurde. Die Nominierungen in dieser Kategorie haben sich seit ~ 1970 gehäuft. Directing als größte Kategorie ist seit 1929 bis heute vergeben worden und die Nominierungen sind während dieser Zeit in etwa gleichverteilt. Die beliebte Kategorie “Best Picture” gibt es erst in etwa seit 1960, während Oscars in der Kategorie Art Direction seit ~2005 nicht mehr vergeben werden.

Nominierte

```r
most_freq_names <- count(oscars_tbl, name) %>% arrange(desc(n)) %>% top_n(10, n)

most_freq_names
```

```
## # A tibble: 11 x 2
##    name                             n
##    <chr>                        <int>
##  1 Metro-Goldwyn-Mayer             64
##  2 Walt Disney, Producer           59
##  3 John Williams                   44
##  4 Warner Bros.                    43
##  5 France                          37
##  6 Alfred Newman                   34
##  7 Italy                           27
##  8 Paramount                       25
##  9 Edith Head                      22
## 10 Gordon Hollingshead, Producer   22
## 11 RKO Radio                       22
```

```r
top_names_filtered <- filter(oscars_tbl, name %in% most_freq_names$name)

ggplot(top_names_filtered, aes(y = fct_rev(fct_infreq(name)), fill = winner)) +
  geom_bar() + labs(title = "Personen, Organisationen oder Länder mit min. 22 Nominierungen", x = "Anzal
                    y = "Nominierte") +
  scale_fill_discrete(guide = guide_legend(reverse=TRUE), name = "Auszeichnung", labels = c("Nein", "Ja
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_text(hjust = 0.5))
```



Metro-Goldwyn-Mayer wurde mit rund 64 Mal am meisten nominiert. Die Verteilung zwischen Nominierung und Auszeichnung scheint nicht gleichmäßig verteilt - Walt Disney, Producer hat rund doppelt so viele Auszeichnungen wie Metro-Goldwyn-Meyer bei 5 Nominierungen weniger (59).

Nach Kategorie "DIRECTING"

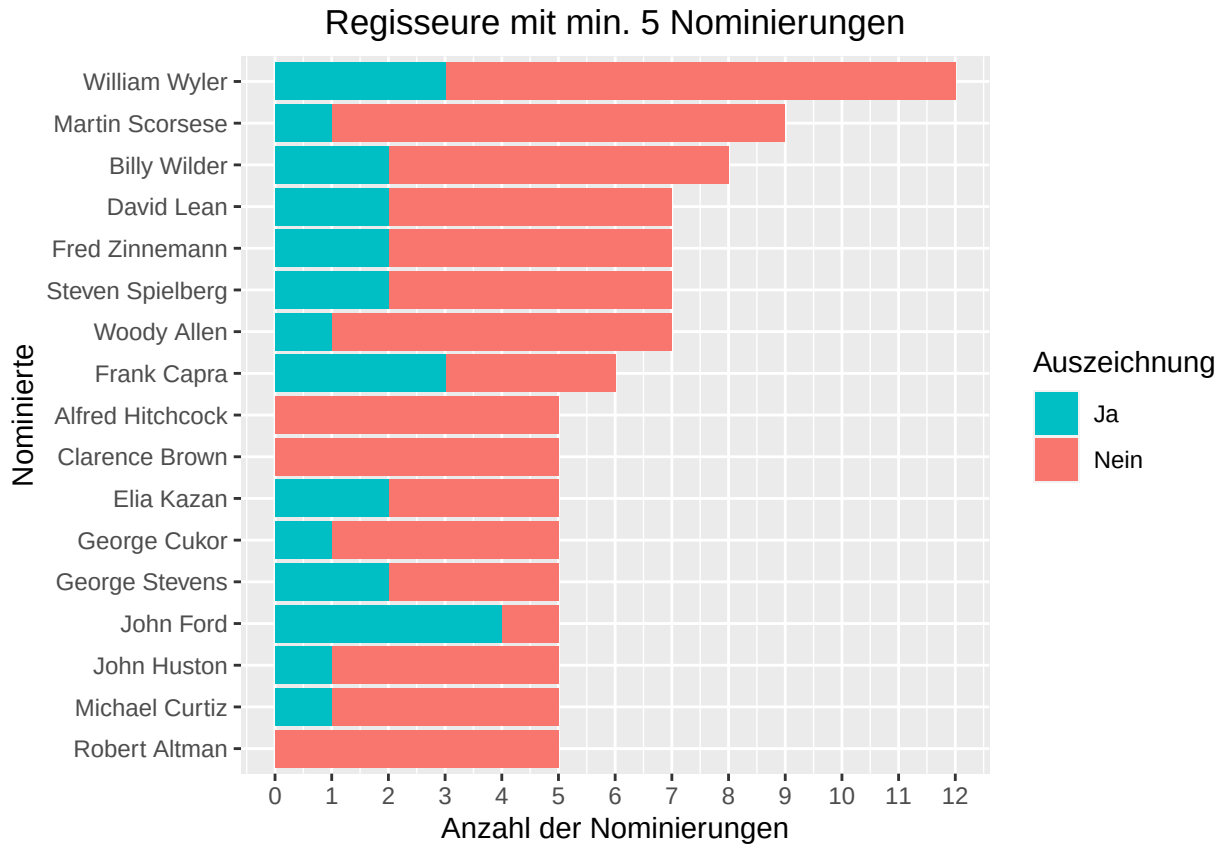
Sortiert nach der häufigsten Kategorie "Directing"

```
top_directors_nom <- filter(oscars_tbl, category == "DIRECTING") %>% count(name) %>% arrange(desc(n)) %
top_directors_nom
```

```
## # A tibble: 17 x 2
##    name                n
##    <chr>           <int>
##  1 William Wyler      12
##  2 Martin Scorsese     9
##  3 Billy Wilder        8
##  4 David Lean          7
##  5 Fred Zinnemann      7
##  6 Steven Spielberg    7
##  7 Woody Allen         7
##  8 Frank Capra         6
##  9 Alfred Hitchcock    5
## 10 Clarence Brown      5
```

```
## 11 Elia Kazan           5
## 12 George Cukor         5
## 13 George Stevens       5
## 14 John Ford            5
## 15 John Huston          5
## 16 Michael Curtiz       5
## 17 Robert Altman        5
```

```
top_directors_filtered <- filter(oscars_tbl, category == "DIRECTING") %>%
  filter(name %in% top_directors_nom$name)

ggplot(top_directors_filtered, aes(y = fct_rev(fct_infreq(name)), fill = winner)) +
  geom_bar() + labs(title = "Regisseure mit min. 5 Nominierungen", x = "Anzahl der Nominierungen",
                    y = "Nominierte") +
  scale_fill_discrete(guide = guide_legend(reverse=TRUE), name = "Auszeichnung", labels = c("Nein", "Ja
  scale_x_continuous(breaks=c(0:12)) +
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_text(hjust = 0.5))
```




```
sum(top_directors_nom$n)/biggest_categories$n[2] # Top Regisseure prozentueller Anteil an Directing Nom
```

```
## [1] 0.2405345
```

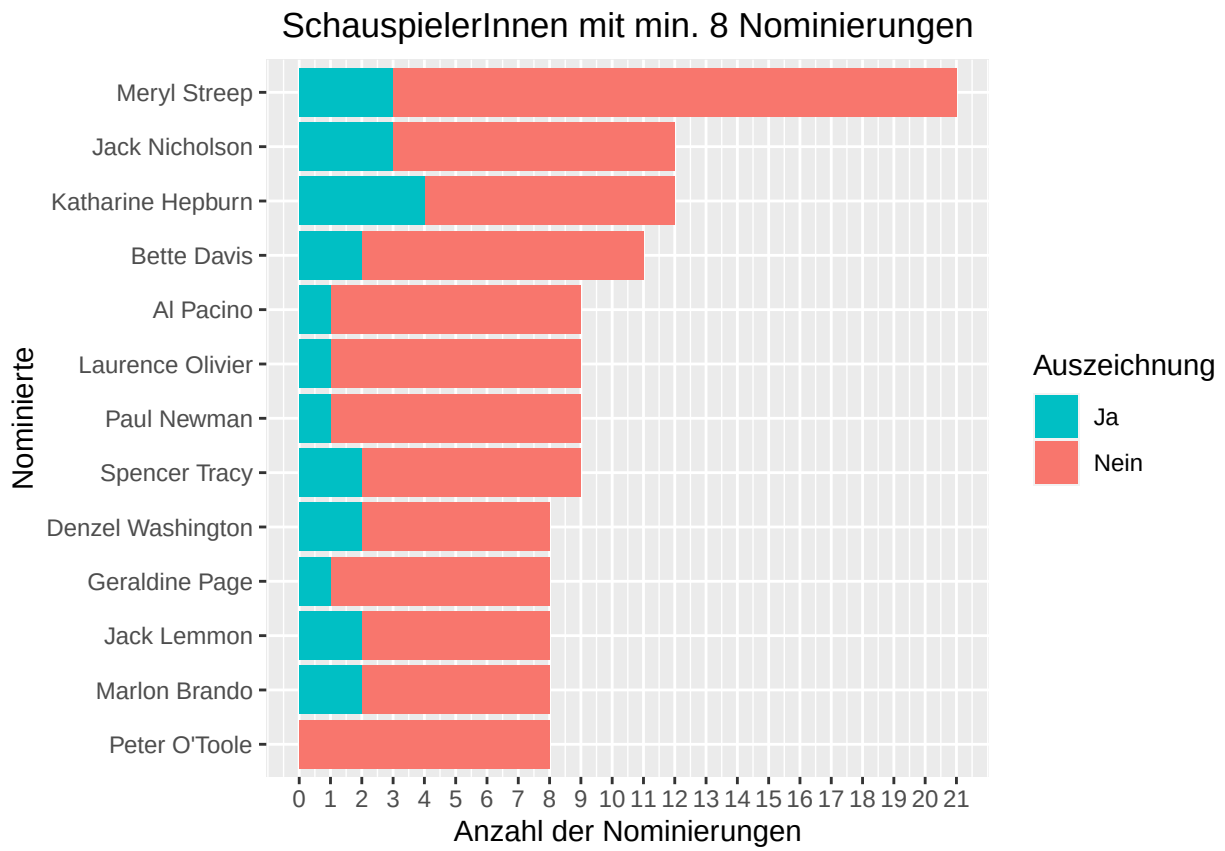
17 Regisseure haben zumindest 5 Nominierungen. William Wyler hat mit 12 die meisten Nominierungen, John Ford jedoch mit 4 Auszeichnungen aus 5 Nominierungen die meisten Preise in der Kategorie.

Nach Kategorie Actor/Actress (beliebig)

```
top_actors_n <- filter(oscars_tbl, grepl('ACTOR*|ACTRESS*', category)) %>% count(name) %>% arrange(desc
top_actors_n
```

```
## # A tibble: 13 x 2
##    name                  n
##    <chr>             <int>
##  1 Meryl Streep         21
##  2 Jack Nicholson       12
##  3 Katharine Hepburn    12
##  4 Bette Davis          11
##  5 Al Pacino             9
##  6 Laurence Olivier      9
##  7 Paul Newman           9
##  8 Spencer Tracy         9
##  9 Denzel Washington     8
## 10 Geraldine Page        8
## 11 Jack Lemmon           8
## 12 Marlon Brando         8
## 13 Peter O'Toole         8
```

```
top_actors_n_filtered <- filter(oscars_tbl, grepl('ACTOR*|ACTRESS*', category)) %>%
  filter(name %in% top_actors_n$name)

ggplot(top_actors_n_filtered, aes(y = fct_rev(fct_infreq(name)), fill = winner)) +
  geom_bar() + labs(title = "SchauspielerInnen mit min. 8 Nominierungen", x = "Anzahl der Nominierungen
                    y = "Nominierte") +
  scale_fill_discrete(guide = guide_legend(reverse=TRUE), name = "Auszeichnung", labels = c("Nein", "Ja
  scale_x_continuous(breaks=c(0:21)) +
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_text(hjust = 0.5))
```



Meryl Streep hat mit Abstand die meisten Nominierungen (21) aller SchauspielerInnnen, fast doppelt so viel wie Jack Nicholson und Katharine Hepburn (12) mit den zweitmeisten Nominierungen. Katharine Hepburn hat jedoch die meisten Preise in den Schauspiel-Kategorien (4) gewonnen.

Filme

```
most_freq_films <- oscars_tbl %>% drop_na() %>%  count(film,year_film) %>% arrange(desc(n)) %>% top_n(5
most_freq_films
```

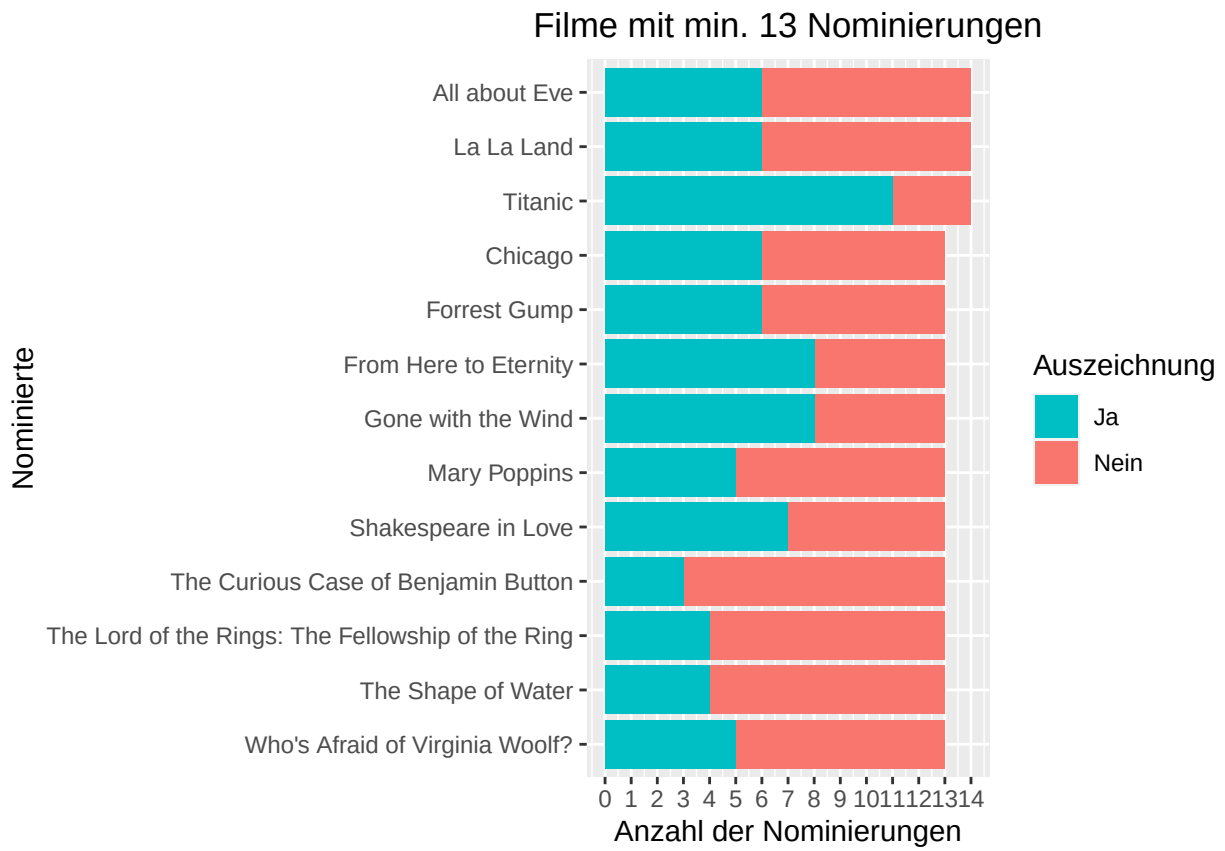
```
## # A tibble: 13 x 3
##    film                                                year_film     n
##    <chr>                                                   <int> <int>
##  1 All about Eve                                            1950    14
##  2 La La Land                                               2016    14
##  3 Titanic                                                  1997    14
##  4 Chicago                                                  2002    13
##  5 Forrest Gump                                             1994    13
##  6 From Here to Eternity                                    1953    13
##  7 Gone with the Wind                                       1939    13
##  8 Mary Poppins                                             1964    13
##  9 Shakespeare in Love                                      1998    13
## 10 The Curious Case of Benjamin Button                      2008    13
## 11 The Lord of the Rings: The Fellowship of the Ring        2001    13
```

```
## 12 The Shape of Water                                    2017    13
## 13 Who's Afraid of Virginia Woolf?                       1966    13
```

```
top_films_filtered <- oscars_tbl %>% filter(film %in% most_freq_films$film) %>% filter(!(film == 'Titan
top_films_filtered
```

```
## # A tibble: 172 x 7
##    year_film year_ceremony ceremony category       name          film     winner
##        <int>         <int>    <int> <fct>          <chr>         <chr>    <fct>
##  1      1939          1940       12 ACTOR          Clark Gable   Gone wi~ False
##  2      1939          1940       12 ACTRESS        Vivien Leigh  Gone wi~ True
##  3      1939          1940       12 ACTRESS IN A ~ Olivia de Ha~ Gone wi~ False
##  4      1939          1940       12 ACTRESS IN A ~ Hattie McDan~ Gone wi~ True
##  5      1939          1940       12 ART DIRECTION  Lyle Wheeler  Gone wi~ True
##  6      1939          1940       12 CINEMATOGRAPH~ Ernest Halle~ Gone wi~ True
##  7      1939          1940       12 DIRECTING      Victor Flemi~ Gone wi~ True
##  8      1939          1940       12 FILM EDITING   Hal C. Kern,~ Gone wi~ True
##  9      1939          1940       12 MUSIC (Origin~ Max Steiner   Gone wi~ False
## 10      1939          1940       12 OUTSTANDING P~ Selznick Int~ Gone wi~ True
## # ... with 162 more rows
```

```
ggplot(top_films_filtered, aes(y = fct_rev(fct_infreq(film)), fill = winner)) +
  geom_bar() + labs(title = "Filme mit min. 13 Nominierungen", x = "Anzahl der Nominierungen",
                    y = "Nominierte") +
  scale_fill_discrete(guide = guide_legend(reverse=TRUE), name = "Auszeichnung", labels = c("Nein", "Ja
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=c(0:14))
```



Die drei meist-nominierten Filme sind “All about Eve”, “La la Land” und “Titanic” (alle 14 Nominierungen), wobei Titanic mit 11 Mal am meisten ausgezeichnet wurde.

Auszeichnungen

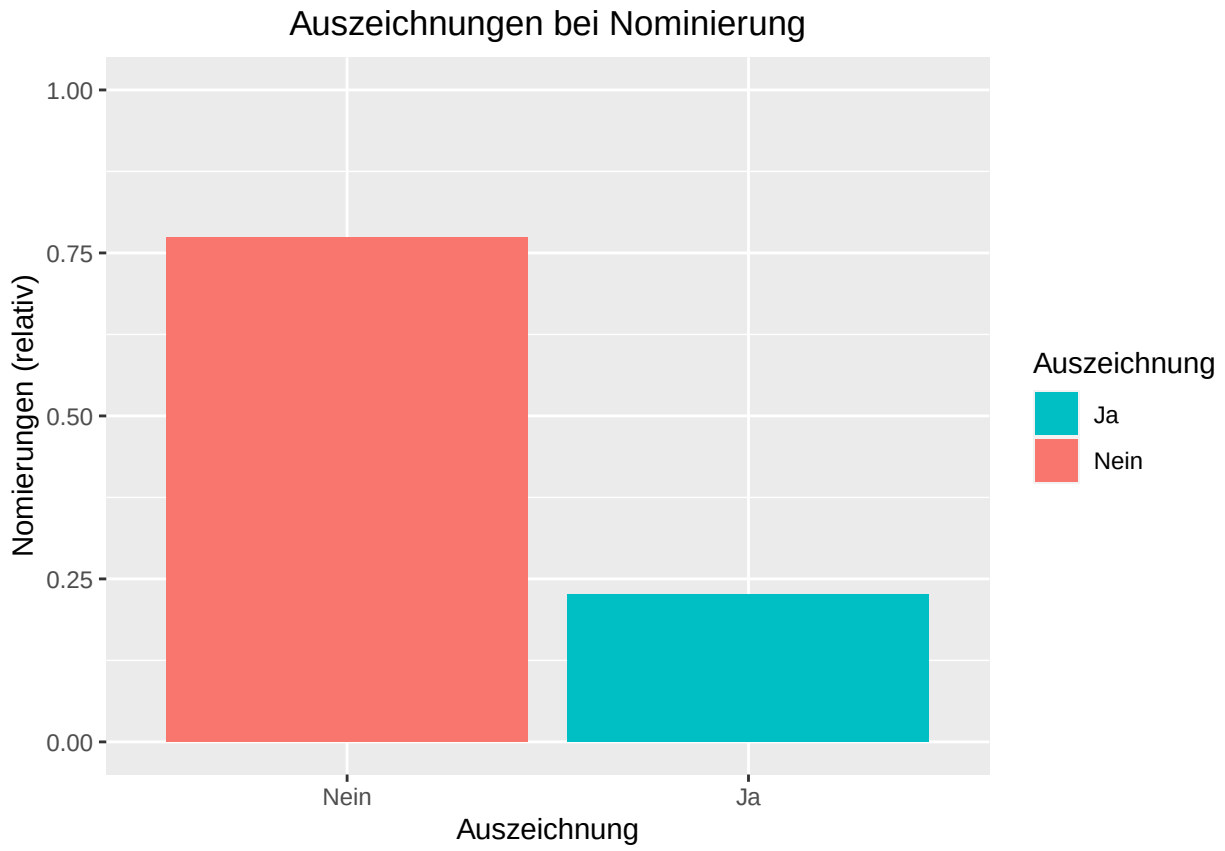
```
abs = table(oscars_tbl$winner)
rel = prop.table(table(oscars_tbl$winner))

tab = rbind(Absolut = abs, Anteile = round(rel, 2))

pander(tab, justify = "right", caption = "Auszeichnungen von Nominierungen", ) # Gibt Tabellen in Markd
```

Table 1: Auszeichnungen von Nominierungen

	False	True
Absolut	8038	2357
Anteile	0.77	0.23

```
ggplot(oscars_tbl, aes(x = winner, y = ..prop.., group = 1, fill = factor(..x..))) +
  geom_bar() +
  scale_fill_discrete(guide = guide_legend(reverse=TRUE), name = "Auszeichnung", labels = c("Nein", "Ja
```

```
  scale_x_discrete(labels = (c("Nein", "Ja"))) +
  scale_y_continuous(limits = c(0,1)) +
  labs(title = "Auszeichnungen bei Nominierung", x = "Auszeichnung",
                    y = "Nomierungen (relativ)") +
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_text(hjust = 0.5))
```



Rund 23% (2357) aller Nominierungen werden mit einem Oscar ausgezeichnet.