```
> library(tidyverse)
> library(dslabs)
> data("movielens")
>
> dim(movielens)
[1] 100004      7
> head(movielens)
  movieId                                    title year
1      31                     Dangerous Minds 1995
2    1029                                Dumbo 1941
3    1061                              Sleepers 1996
4    1129                Escape from New York 1981
5    1172 Cinema Paradiso (Nuovo cinema Paradiso) 1989
6    1263                      Deer Hunter, The 1978
                          genres userId rating  timestamp
1                          Drama      1    2.5 1260759144
2 Animation|Children|Drama|Musical      1    3.0 1260759179
3                       Thriller      1    3.0 1260759182
4 Action|Adventure|Sci-Fi|Thriller      1    2.0 1260759185
5                          Drama      1    4.0 1260759205
6                     Drama|War      1    2.0 1260759151
>
>
>
> movielens %>% group_by(movieId) %>% summarize(n = n(), year = as.character(first(year)))
# A tibble: 9,066 x 3
   movieId     n year
     <int> <int> <chr>
 1       1   247 1995
 2       2   107 1995
 3       3    59 1995
 4       4    13 1995
 5       5    56 1995
 6       6   104 1995
 7       7    53 1995
 8       8     5 1995
 9       9    20 1995
10      10   122 1995
# … with 9,056 more rows
> movielens %>% group_by(movieId) %>% summarize(n = n(), year = as.character(first(year))) %>% ggplot(aes(
year, n)) + geom_boxplot() + coord_trans(y = 'sqrt') + theme(axis.text.x = element_text(angle = 90, hjust
= 1))
>
>
>
> movielens %>% filter(year >= 1993) %>% group_by(movieId) %>% summarize(n = length(rating), title = first
(title) , rate = n / (2018 - as.numeric(first(year))), rating = mean(rating)) %>% arrange(desc(rate))
# A tibble: 5,666 x 5
   movieId     n title                                                rate rating
     <int> <int> <chr>                                               <dbl>  <dbl>
 1     356   341 Forrest Gump                                         14.2   4.05
 2   79132   111 Inception                                            13.9   4.05
 3    2571   259 Matrix, The                                          13.6   4.18
 4     296   324 Pulp Fiction                                         13.5   4.26
 5     318   311 Shawshank Redemption, The                            13.0   4.49
 6   58559   121 Dark Knight, The                                     12.1   4.24
 7    4993   200 Lord of the Rings: The Fellowship of the Ring, The   11.8   4.18
 8    5952   188 Lord of the Rings: The Two Towers, The               11.8   4.06
 9    7153   176 Lord of the Rings: The Return of the King, The       11.7   4.13
10    2858   220 American Beauty                                      11.6   4.24
# … with 5,656 more rows
> movielens %>% filter(year >= 1993) %>% group_by(movieId) %>% summarize(n = length(rating), title = first
(title) , rate = n / (2018 - as.numeric(first(year))), rating = mean(rating)) %>% arrange(desc(rate)) %>%
top_n(25, rate) %>% arrange(desc(rate))
# A tibble: 25 x 5
```

```
   movieId     n title                                                  rate rating
     <int> <int> <chr>                                                 <dbl>  <dbl>
 1     356   341 Forrest Gump                                          14.2   4.05
 2   79132   111 Inception                                             13.9   4.05
 3    2571   259 Matrix, The                                           13.6   4.18
 4     296   324 Pulp Fiction                                          13.5   4.26
 5     318   311 Shawshank Redemption, The                             13.0   4.49
 6   58559   121 Dark Knight, The                                      12.1   4.24
 7    4993   200 Lord of the Rings: The Fellowship of the Ring, The    11.8   4.18
 8    5952   188 Lord of the Rings: The Two Towers, The                11.8   4.06
 9    7153   176 Lord of the Rings: The Return of the King, The        11.7   4.13
10    2858   220 American Beauty                                       11.6   4.24
# … with 15 more rows
>
>
>
> movielens %>% filter(year >= 1993) %>% group_by(movieId) %>% summarize(n = length(rating), title = first
(title) , rate = n / (2018 - as.numeric(first(year))), rating = mean(rating)) %>% arrange(desc(rate)) %>%
ggplot(aes(rate, rating)) + geom_point() + geom_smooth()
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
>
>
>
> library(lubridate)
> movielens <- mutate(movielens, date = as_datetime(timestamp))
> head(movielens)
  movieId                                    title year
1      31                          Dangerous Minds 1995
2    1029                                   Dumbo 1941
3    1061                                 Sleepers 1996
4    1129                     Escape from New York 1981
5    1172 Cinema Paradiso (Nuovo cinema Paradiso) 1989
6    1263                          Deer Hunter, The 1978
                            genres userId rating  timestamp                date
1                            Drama      1    2.5 1260759144 2009-12-14 02:52:24
2 Animation|Children|Drama|Musical      1    3.0 1260759179 2009-12-14 02:52:59
3                         Thriller      1    3.0 1260759182 2009-12-14 02:53:02
4 Action|Adventure|Sci-Fi|Thriller      1    2.0 1260759185 2009-12-14 02:53:05
5                            Drama      1    4.0 1260759205 2009-12-14 02:53:25
6                       Drama|War      1    2.0 1260759151 2009-12-14 02:52:31
>
>
>
> movielens %>% mutate(date = round_date(date, unit = 'week')) %>% group_by(date)
# A tibble: 100,004 x 8
# Groups:   date [988]
   movieId title      year genres    userId rating timestamp date
     <int> <chr>     <int> <fct>      <int>  <dbl>     <int> <dttm>
 1      31 Dangerou… 1995 Drama          1    2.5    1.26e9 2009-12-13 00:00:00
 2    1029 Dumbo     1941 Animatio…      1    3      1.26e9 2009-12-13 00:00:00
 3    1061 Sleepers  1996 Thriller       1    3      1.26e9 2009-12-13 00:00:00
 4    1129 Escape f… 1981 Action|A…      1    2      1.26e9 2009-12-13 00:00:00
 5    1172 Cinema P… 1989 Drama          1    4      1.26e9 2009-12-13 00:00:00
 6    1263 Deer Hun… 1978 Drama|War      1    2      1.26e9 2009-12-13 00:00:00
 7    1287 Ben-Hur   1959 Action|A…      1    2      1.26e9 2009-12-13 00:00:00
 8    1293 Gandhi    1982 Drama          1    2      1.26e9 2009-12-13 00:00:00
 9    1339 Dracula … 1992 Fantasy|…      1    3.5    1.26e9 2009-12-13 00:00:00
10    1343 Cape Fear 1991 Thriller       1    2      1.26e9 2009-12-13 00:00:00
# … with 99,994 more rows
> movielens %>% mutate(date = round_date(date, unit = 'week')) %>% group_by(date) %>% summarize(rating = m
ean(rating))
# A tibble: 988 x 2
   date                rating
   <dttm>               <dbl>
 1 1995-01-08 00:00:00   3.67
```

```
 2 1996-03-31 00:00:00   4.24
 3 1996-04-07 00:00:00   4
 4 1996-04-14 00:00:00   3.83
 5 1996-04-21 00:00:00   4.23
 6 1996-04-28 00:00:00   4
 7 1996-05-05 00:00:00   3.56
 8 1996-05-12 00:00:00   3.51
 9 1996-05-19 00:00:00   3.45
10 1996-05-26 00:00:00   3.73
# … with 978 more rows
> movielens %>% mutate(date = round_date(date, unit = 'week')) %>% group_by(date) %>% summarize(rating = m
ean(rating)) %>%  ggplot(aes(date, rating)) + geom_point() + geom_smooth()
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
>
>
>
> movielens %>% group_by(genres) %>% summarize(n = n(), avg = mean(rating), se = sd(rating)/sqrt(n()))
# A tibble: 901 x 4
   genres                                             n   avg     se
   <fct>                                          <int> <dbl>  <dbl>
 1 (no genres listed)                                18  3.78 0.311
 2 Action                                           143  2.88 0.0872
 3 Action|Adventure                                 540  3.79 0.0449
 4 Action|Adventure|Animation                        23  3.98 0.115
 5 Action|Adventure|Animation|Children               15  3.57 0.258
 6 Action|Adventure|Animation|Children|Comedy        147  3.83 0.0720
 7 Action|Adventure|Animation|Children|Comedy|Fantasy  14  3.93 0.267
 8 Action|Adventure|Animation|Children|Comedy|IMAX     14  3.46 0.270
 9 Action|Adventure|Animation|Children|Comedy|Romance   6  3.33 0.511
10 Action|Adventure|Animation|Children|Comedy|Sci-Fi   14  2.89 0.278
# … with 891 more rows
> movielens %>% group_by(genres) %>% summarize(n = n(), avg = mean(rating), se = sd(rating)/sqrt(n())) %>%
 filter(n >= 1000)
# A tibble: 18 x 4
   genres                              n   avg     se
   <fct>                           <int> <dbl>  <dbl>
 1 Action|Adventure|Sci-Fi          2146  3.58 0.0240
 2 Action|Adventure|Sci-Fi|Thriller  1453  3.49 0.0265
 3 Action|Adventure|Thriller         1413  3.38 0.0258
 4 Action|Crime|Thriller             1441  3.48 0.0258
 5 Action|Drama|War                  1019  3.84 0.0316
 6 Action|Sci-Fi|Thriller            1065  3.61 0.0328
 7 Comedy                            6748  3.27 0.0135
 8 Comedy|Crime                      1096  3.41 0.0351
 9 Comedy|Drama                      3272  3.63 0.0174
10 Comedy|Drama|Romance              3204  3.62 0.0175
11 Comedy|Romance                    3973  3.37 0.0167
12 Crime|Drama                       2367  4.01 0.0188
13 Crime|Drama|Thriller              1091  3.71 0.0290
14 Documentary                       1154  3.86 0.0273
15 Drama                             7757  3.71 0.0111
16 Drama|Romance                     3462  3.66 0.0175
17 Drama|Thriller                    1402  3.46 0.0260
18 Drama|War                         1167  4.01 0.0277
> movielens %>% group_by(genres) %>% summarize(n = n(), avg = mean(rating), se = sd(rating)/sqrt(n())) %>%
 filter(n >= 1000) %>%  mutate(genres = reorder(genres, avg))
# A tibble: 18 x 4
   genres                              n   avg     se
   <fct>                           <int> <dbl>  <dbl>
 1 Action|Adventure|Sci-Fi          2146  3.58 0.0240
 2 Action|Adventure|Sci-Fi|Thriller  1453  3.49 0.0265
 3 Action|Adventure|Thriller         1413  3.38 0.0258
 4 Action|Crime|Thriller             1441  3.48 0.0258
 5 Action|Drama|War                  1019  3.84 0.0316
 6 Action|Sci-Fi|Thriller            1065  3.61 0.0328
```

```
 7 Comedy                            6748  3.27 0.0135
 8 Comedy|Crime                      1096  3.41 0.0351
 9 Comedy|Drama                      3272  3.63 0.0174
10 Comedy|Drama|Romance              3204  3.62 0.0175
11 Comedy|Romance                    3973  3.37 0.0167
12 Crime|Drama                       2367  4.01 0.0188
13 Crime|Drama|Thriller              1091  3.71 0.0290
14 Documentary                       1154  3.86 0.0273
15 Drama                             7757  3.71 0.0111
16 Drama|Romance                     3462  3.66 0.0175
17 Drama|Thriller                    1402  3.46 0.0260
18 Drama|War                         1167  4.01 0.0277
> movielens %>% group_by(genres) %>% summarize(n = n(), avg = mean(rating), se = sd(rating)/sqrt(n())) %>%
 filter(n >= 1000) %>% mutate(genres = reorder(genres, avg)) %>% ggplot(aes(x = genres, y = avg, ymin = av
g - 2*se, ymax = avg + 2*se)) + geom_point() + geom_errorbar() + theme(axis.text.x = element_text(angle =
90, hjust = 1))
>
```