Understanding the data

Emilia Pankowska

4 listopada 2016

Loading the data

The "data.csv" should be in working directory.

Empty strings are replaced with NA

```
df = read.table("data.csv", sep =",", head = TRUE, na.strings = "")
```

Getting look at the data

```
head(df)
```

```
##   learner_id country in_course     unit avg_score completion inv_rate
## 1         39      PL         t        1     0.200      0.100    0.000
## 2         39      PL         t        3     1.000      0.087    0.000
## 3         39      PL         t REVIEW 2     0.100      0.273    0.000
## 4         41      PL         t        1     0.877      0.350    0.381
## 5         41      PL         t        3     0.000      0.087    0.000
## 6         41      PL         t        5     0.000      0.087    1.000
```

```
str(df)
```

```
## 'data.frame':    81432 obs. of  7 variables:
##  $ learner_id: int  39 39 39 41 41 41 41 802 802 802 ...
##  $ country   : Factor w/ 86 levels "AD","AF","AG",..: 61 61 61 61 61 61 61 NA NA NA ...
##  $ in_course : Factor w/ 2 levels "f","t": 2 2 2 2 2 2 2 2 2 2 ...
##  $ unit      : Factor w/ 17 levels "1","10","11",..: 1 6 14 1 6 8 14 1 5 6 ...
##  $ avg_score : num  0.2 1 0.1 0.877 0 0 0 0.954 0.877 0.883 ...
##  $ completion: num  0.1 0.087 0.273 0.35 0.087 0.087 0.182 1 1 0.739 ...
##  $ inv_rate  : num  0 0 0 0.381 0 1 0 0 0 0 ...
```

```
summary(df)
```

```
##    learner_id         country      in_course      unit
##  Min.   :     39   TR     :52757   f: 6100   1      :11435
##  1st Qu.:1044868   ES     : 8409   t:75332   2      : 9273
##  Median :1092952   PL     : 5925             3      : 8163
##  Mean   :1137210   CO     : 5537             4      : 7136
##  3rd Qu.:1244376   IT     : 1779             5      : 6448
##  Max.   :1390141   (Other): 7017             6      : 5916
##                    NA's   :    8             (Other):33061
##    avg_score        completion        inv_rate
```

```
##  Min.   :0.0000   Min.   :0.0250   Min.   :0.00000
##  1st Qu.:0.7170   1st Qu.:0.7000   1st Qu.:0.00000
##  Median :0.8400   Median :0.9620   Median :0.00000
##  Mean   :0.7921   Mean   :0.8098   Mean   :0.04416
##  3rd Qu.:0.9300   3rd Qu.:1.0000   3rd Qu.:0.02300
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##  NA's   :2
```

Cleaning the data

Checking *strange* data

The result of *summary* function shows that:

1. there are some missing data in *country* and *avg_score* variables
2. Variables *avg_score*, *completion* and *inv_rate* are between 0 and 1 – there is no outlier values or typos.

The values of factor variables can be listed to check if there is anything unexpected.

```
apply(df[, 2:4], 2, function(x) sort(table(x)))
```

```
## $country
## x
##    BD    PT    ZW    AT    CA    PM    CK    IR    SV    SZ    AN    AO
##     1     1     1     2     2     2     3     3     3     3     4     4
##    IQ    PH    AI    DZ    HK    LY    GM    ML    CD    CL    VA    BL
##     4     4     5     5     5     5     6     6     7     8     9    10
##    HR    MA    VE    AG    LV    CR    TW    VN    MD    MK    TC    AX
##    10    10    10    11    11    12    12    12    13    13    13    14
##    KW    PS    TN    AL    CY    SK    SO    GR    ID    AF    BR    AS
##    14    16    16    17    17    17    18    23    23    24    24    26
##    YE    EC    TH    AZ    DE    AQ    BG    TM    KR    LT    JP    FR
##    26    29    29    35    37    42    46    48    58    68    73    78
##    AD    AR    NZ    US    GB    SA    BY    QU    UA    HU    BE    RO
##    83    89    94    94    96   111   121   134   159   167   180   218
##    CN    TL    RU    AU    CZ    MX    CH    NL    OM    IT    CO    PL
##   233   260   276   432   464   517   554   831   856  1779  5537  5925
##    ES    TR
##  8409 52757
##
## $in_course
## x
##     f     t
##  6100 75332
##
## $unit
## x
##       REVIEW 4             12             11             10       REVIEW 3
##            483            938           1165           1757           2385
## VIDEO PODCASTS              9       REVIEW 2              8              7
##           3367           3610           3827           4834           5300
```

```
##       REVIEW 1              6              5              4              3
##           5395           5916           6448           7136           8163
##              2              1
##           9273          11435
```

All values make senes. There is a

Dealing with missing data

Because there is only few missing cases, it is possible to print them:

```
df[!complete.cases(df),]
```

```
##       learner_id country in_course           unit avg_score completion
## 8            802    <NA>         t              1     0.954      1.000
## 9            802    <NA>         t              2     0.877      1.000
## 10           802    <NA>         t              3     0.883      0.739
## 11           802    <NA>         t       REVIEW 1     0.904      1.000
## 12           802    <NA>         t VIDEO PODCASTS     0.955      0.175
## 15          5811    <NA>         t              1     0.967      0.300
## 16          5811    <NA>         t              2     0.611      0.154
## 17          5811    <NA>         t              8     0.557      0.450
## 11441    1037376      TR         f             10        NA      0.143
## 54050    1187879      CO         f              1        NA      0.150
##       inv_rate
## 8        0.000
## 9        0.000
## 10       0.000
## 11       0.141
## 12       0.000
## 15       0.000
## 16       0.000
## 17       0.000
## 11441    0.000
## 54050    0.000
```

There are two learners with missing country Checking if it's possible to get this info from data:

```
miss_country <- unique(df[is.na(df$country), 1])
unique(df[df$learner_id %in% miss_country,1:2 ])
```

```
##    learner_id country
## 8         802    <NA>
## 15       5811    <NA>
```

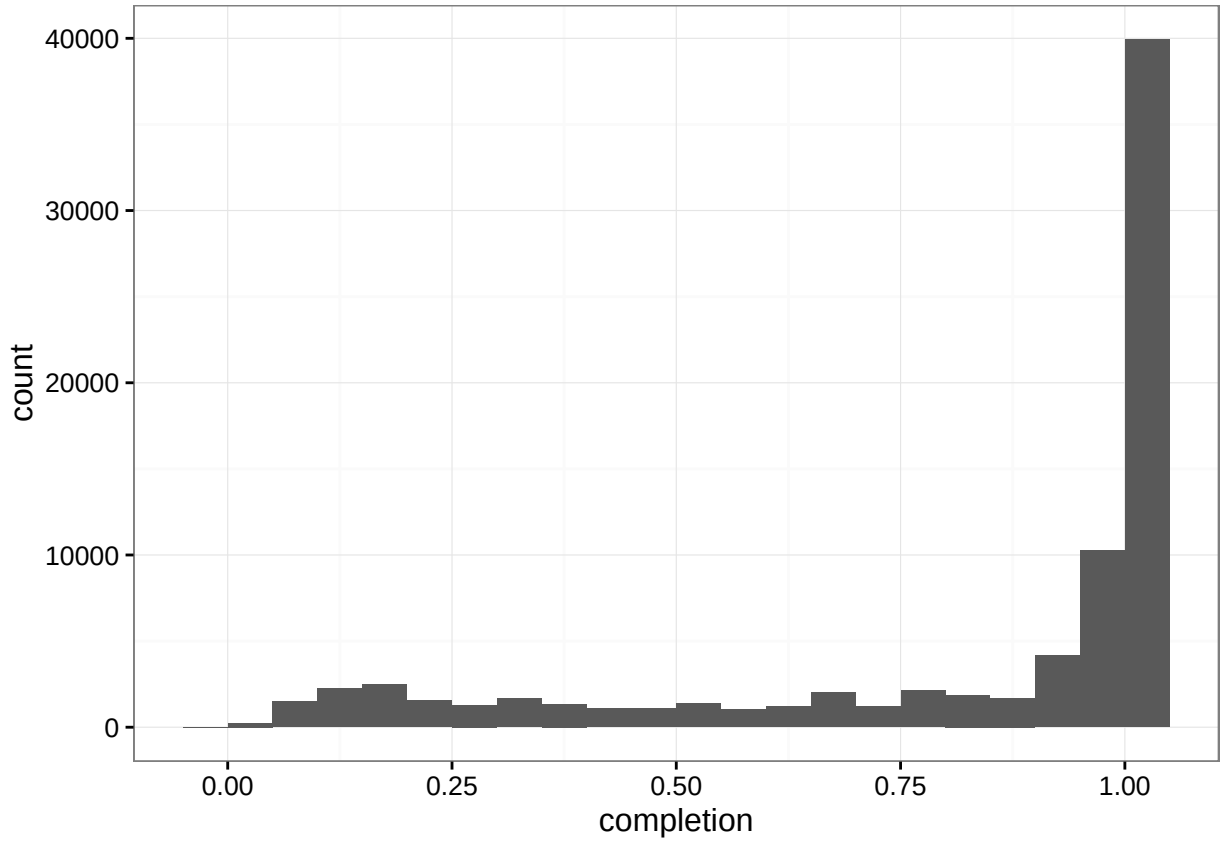
Because there are only ten (of over 80K) cases with missing values, there is no point of replacing them and they can be ignored.

Removing missing data:

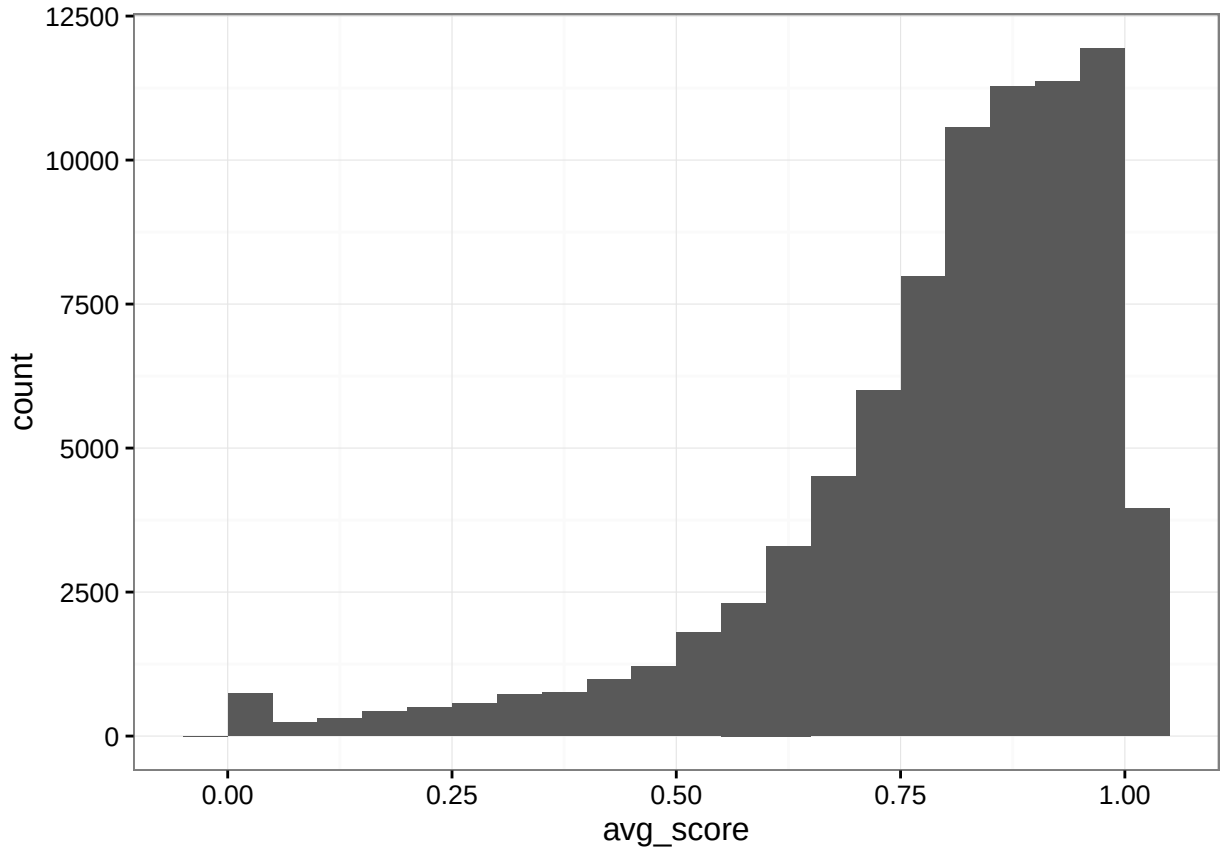
```r
df <- df[complete.cases(df), ]
```

Some plots

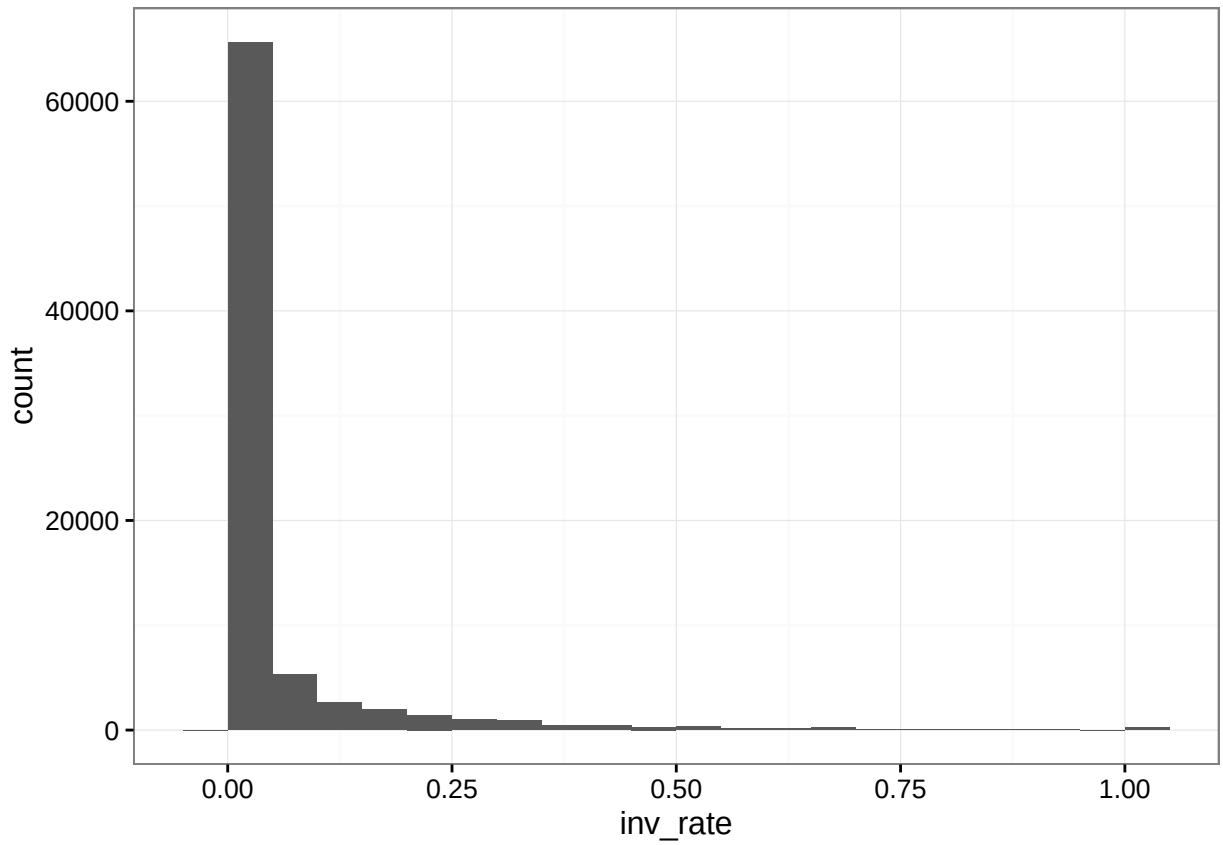
```r
ggplot(df, aes(x = completion)) + geom_histogram(binwidth = 0.05)
```



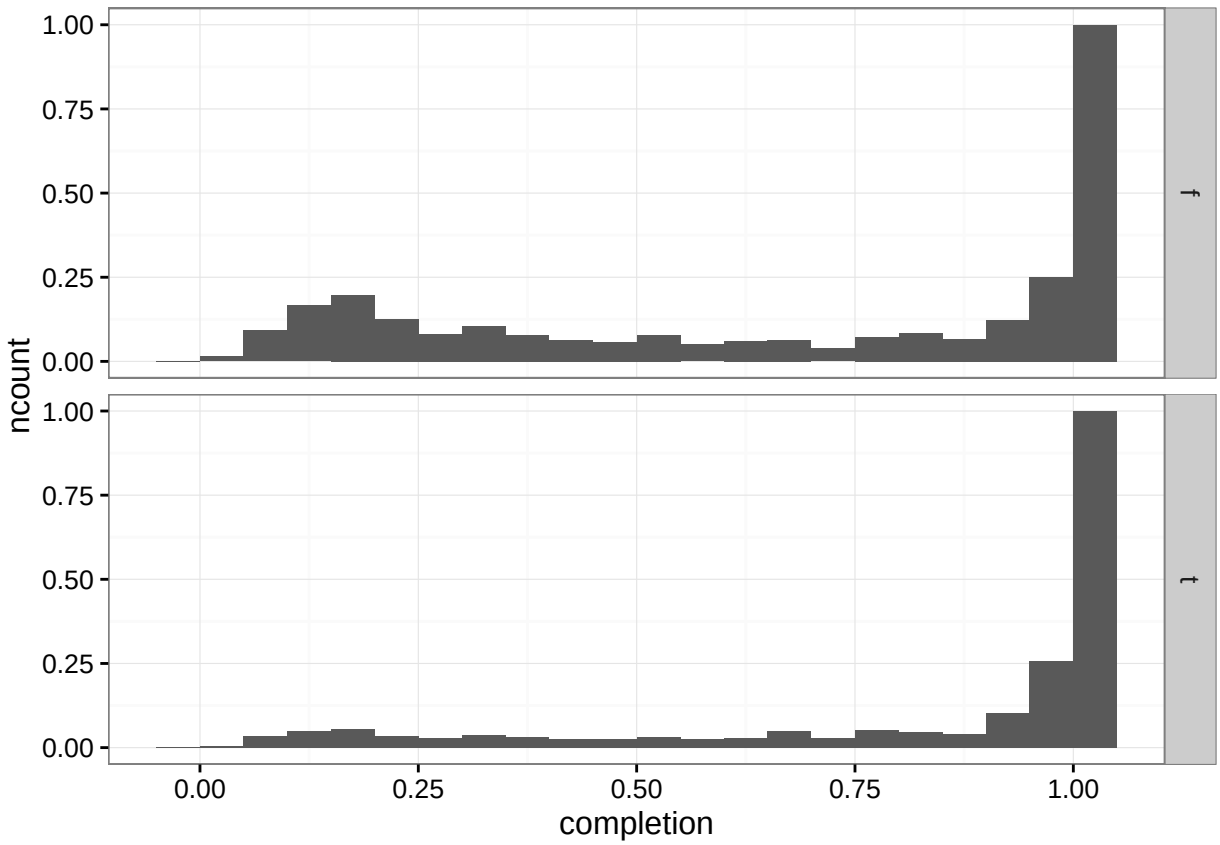

```r
ggplot(df, aes(x = avg_score)) + geom_histogram(binwidth = 0.05)
```



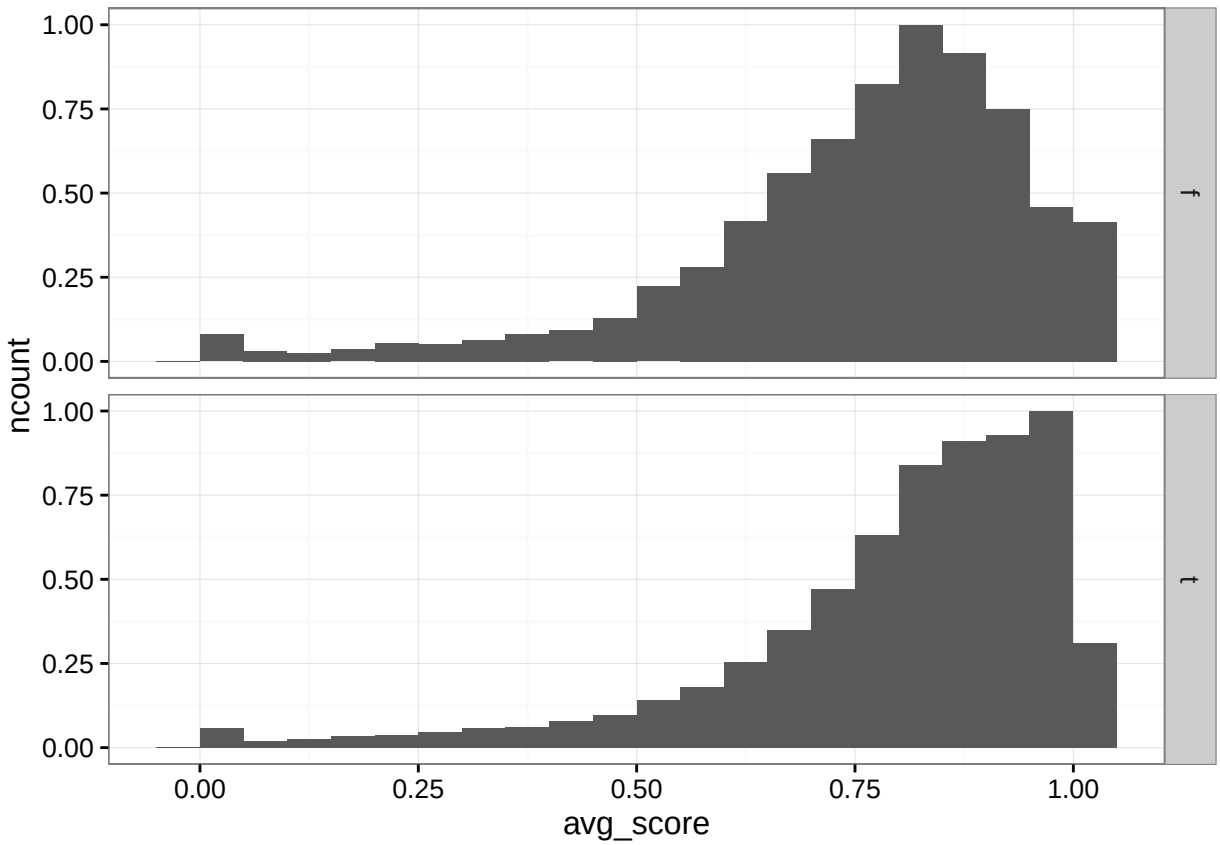

```
ggplot(df, aes(x = inv_rate)) + geom_histogram(binwidth = 0.05)
```



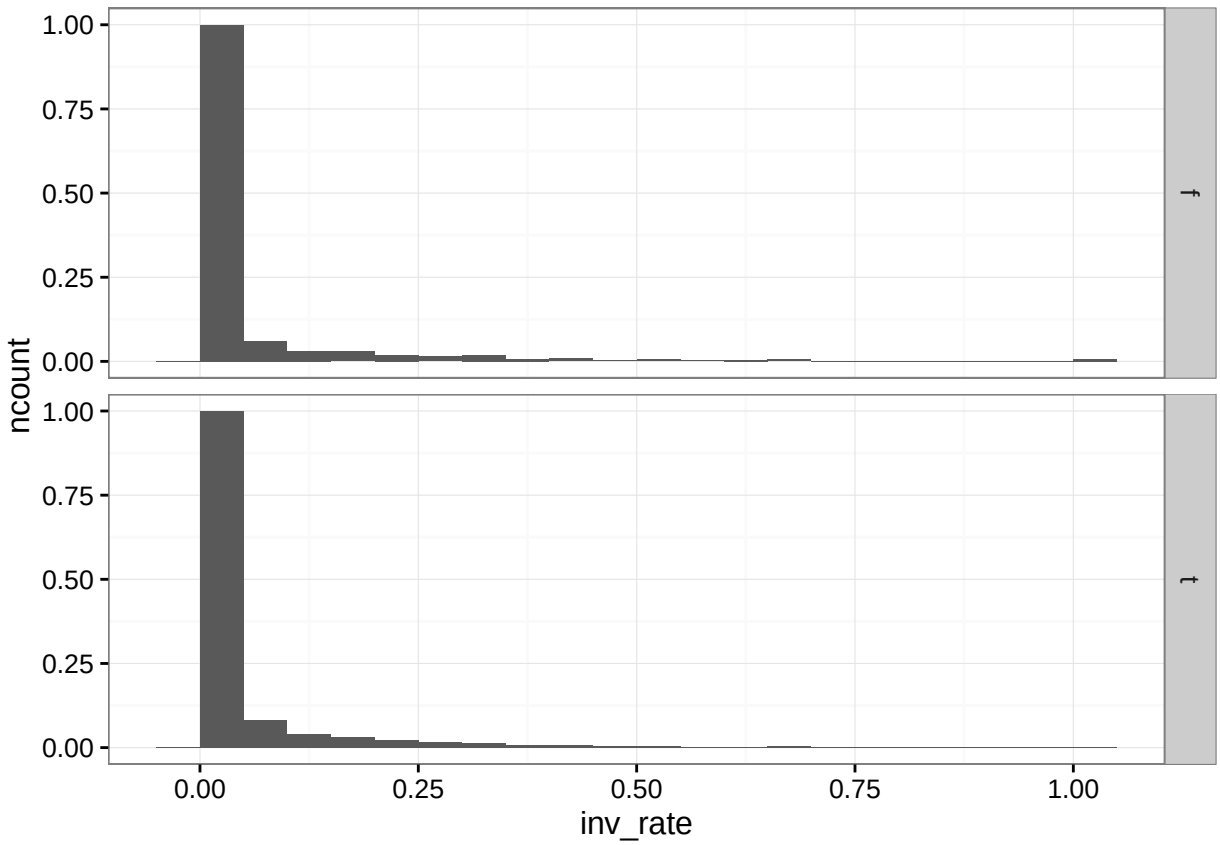

```
ggplot(df, aes(x = completion)) + geom_histogram(binwidth = 0.05, aes(y=..ncount..)) + facet_grid(in_cou
```




```
ggplot(df, aes(x = avg_score)) + geom_histogram(binwidth = 0.05, aes(y=..ncount..)) + facet_grid(in_cou
```




```
ggplot(df, aes(x = inv_rate)) + geom_histogram(binwidth = 0.05, aes(y=..ncount..)) + facet_grid(in_cours
```


1.00
0.75
0.50
0.25
0.00
ncount
f
t
0.00
0.25
0.50
0.75
1.00
inv_rate