# R：data.table 包 vs. Pandas

2023 年 11 月 14 日

[3]:
```
# 调用 R
import rpy2.ipython
%load_ext rpy2.ipython
```

[1]:
```
import pandas as pd
pd.__version__
```

[1]:
```
'1.3.5'
```

[4]:
```
%%R
library(data.table)
packageVersion('data.table')
```

```
R[write to console]: data.table 1.14.8 使用 1
线程 (请参阅?getDTthreads)。最新的消息: r-datatable.com

R[write to console]: **********
用中文运行 data.table。软件包只提供英语支持。当在在线搜索帮助时，也要确保检查英语错误
信息。这个可以通过查看软件包源文件中的 po/R-zh_CN.po 和
po/zh_CN.po 文件获得，这个文件可以并排找到母语和英语错误信息。
**********

R[write to console]: **********
data.table 的安装未检测到 OpenMP 支持。在单线程模式下应该仍能运行
此设备为 Mac。请阅读 https://mac.r-project.org/openmp/。请与 Apple 公司联系以获取支
持。查看
r-datatable.com 以获取更新，并参阅我们的 Mac
设备说明: https://github.com/Rdatatable/data.table/wiki/Installation 在 Mac
```

上出现相关安装问题的报告已数年之久，需要指出的是在 Windows 或 Linux 平台上一般不存在类似问题。

**********

```
[1] '1.14.8'
```

### 0.1 数据加载

[6]:
```
url = "https://vincentarelbundock.github.io/Rdatasets/csv/datasets/HairEyeColor.csv"
df = pd.read_csv(url)
df
```

[6]:

| | rownames | Hair | Eye | Sex | Freq |
|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 |
| 1 | 2 | Brown | Brown | Male | 53 |
| 2 | 3 | Red | Brown | Male | 10 |
| 3 | 4 | Blond | Brown | Male | 3 |
| 4 | 5 | Black | Blue | Male | 11 |
| 5 | 6 | Brown | Blue | Male | 50 |
| 6 | 7 | Red | Blue | Male | 10 |
| 7 | 8 | Blond | Blue | Male | 30 |
| 8 | 9 | Black | Hazel | Male | 10 |
| 9 | 10 | Brown | Hazel | Male | 25 |
| 10 | 11 | Red | Hazel | Male | 7 |
| 11 | 12 | Blond | Hazel | Male | 5 |
| 12 | 13 | Black | Green | Male | 3 |
| 13 | 14 | Brown | Green | Male | 15 |
| 14 | 15 | Red | Green | Male | 7 |
| 15 | 16 | Blond | Green | Male | 8 |
| 16 | 17 | Black | Brown | Female | 36 |
| 17 | 18 | Brown | Brown | Female | 66 |
| 18 | 19 | Red | Brown | Female | 16 |
| 19 | 20 | Blond | Brown | Female | 4 |
| 20 | 21 | Black | Blue | Female | 9 |
| 21 | 22 | Brown | Blue | Female | 34 |

```
22        23    Red   Blue  Female     7
23        24  Blond   Blue  Female    64
24        25  Black  Hazel  Female     5
25        26  Brown  Hazel  Female    29
26        27    Red  Hazel  Female     7
27        28  Blond  Hazel  Female     5
28        29  Black  Green  Female     2
29        30  Brown  Green  Female    14
30        31    Red  Green  Female     7
31        32  Blond  Green  Female     8
```

[8]:
```
%%R
url = "https://vincentarelbundock.github.io/Rdatasets/csv/datasets/HairEyeColor.
 ↪csv"
dt = fread(url)
dt
```

```
    rownames  Hair   Eye    Sex Freq
 1:        1 Black Brown   Male   32
 2:        2 Brown Brown   Male   53
 3:        3   Red Brown   Male   10
 4:        4 Blond Brown   Male    3
 5:        5 Black  Blue   Male   11
 6:        6 Brown  Blue   Male   50
 7:        7   Red  Blue   Male   10
 8:        8 Blond  Blue   Male   30
 9:        9 Black Hazel   Male   10
10:       10 Brown Hazel   Male   25
11:       11   Red Hazel   Male    7
12:       12 Blond Hazel   Male    5
13:       13 Black Green   Male    3
14:       14 Brown Green   Male   15
15:       15   Red Green   Male    7
16:       16 Blond Green   Male    8
17:       17 Black Brown Female   36
18:       18 Brown Brown Female   66
19:       19   Red Brown Female   16
```

```
20:       20 Blond Brown Female    4
21:       21 Black  Blue Female    9
22:       22 Brown  Blue Female   34
23:       23   Red  Blue Female    7
24:       24 Blond  Blue Female   64
25:       25 Black Hazel Female    5
26:       26 Brown Hazel Female   29
27:       27   Red Hazel Female    7
28:       28 Blond Hazel Female    5
29:       29 Black Green Female    2
30:       30 Brown Green Female   14
31:       31   Red Green Female    7
32:       32 Blond Green Female    8
    rownames  Hair   Eye    Sex Freq
```

### 0.2 查看数据结构

[9]:
```
# 数据类型
type(df)
df.dtypes
```

[9]:
```
rownames     int64
Hair        object
Eye         object
Sex         object
Freq         int64
dtype: object
```

[10]:
```
%%R
# 数据类型
class(dt)
str(dt)
```

```
Classes ‘data.table’ and 'data.frame':  32 obs. of  5 variables:
 $ rownames: int  1 2 3 4 5 6 7 8 9 10 …
 $ Hair    : chr  "Black" "Brown" "Red" "Blond" …
 $ Eye     : chr  "Brown" "Brown" "Brown" "Brown" …
```

```
$ Sex     : chr  "Male" "Male" "Male" "Male" …
$ Freq    : int  32 53 10 3 11 50 10 30 10 25 …
- attr(*, ".internal.selfref")=<externalptr>
```

[11]:
```
list(df) # 列名
```

[11]:
```
['rownames', 'Hair', 'Eye', 'Sex', 'Freq']
```

[12]:
```
%R names(dt) # 列名
```

[12]:
```
<rpy2.robjects.vectors.StrVector object at 0x7fc79905a448> [RTYPES.STRSXP]
R classes: ('character',)
['rownam…, 'Hair', 'Eye', 'Sex', 'Freq']
```

[13]:
```
# 打印前后几行
df.head(n=3)
df.tail(n=3)
```

[13]:

| | rownames | Hair | Eye | Sex | Freq |
|---|---|---|---|---|---|
| 29 | 30 | Brown | Green | Female | 14 |
| 30 | 31 | Red | Green | Female | 7 |
| 31 | 32 | Blond | Green | Female | 8 |

[15]:
```
%%R
# 打印前后几行
head(dt, n=3)
tail(dt, n=3)
```

```
   rownames  Hair   Eye    Sex Freq
1:       30 Brown Green Female   14
2:       31   Red Green Female    7
3:       32 Blond Green Female    8
```

[16]:
```
# 维度
df.shape
len(df.index)
len(df.columns)
```

[16]:
```
5
```

[17]:
```
%%R
# 维度
dim(dt)
nrow(dt)
ncol(dt)
```

```
[1] 5
```

[18]:
```
df.describe() # 统计描述
```

[18]:

| | rownames | Freq |
|---|---|---|
| count | 32.000000 | 32.000000 |
| mean | 16.500000 | 18.500000 |
| std | 9.380832 | 18.242099 |
| min | 1.000000 | 2.000000 |
| 25% | 8.750000 | 7.000000 |
| 50% | 16.500000 | 10.000000 |
| 75% | 24.250000 | 29.250000 |
| max | 32.000000 | 66.000000 |

[19]:
```
%R summary(dt)  # 统计描述
```

[19]:
```
<rpy2.robjects.vectors.StrMatrix object at 0x7fc7bab17348> [RTYPES.STRSXP]
R classes: ('table',)
['Min.   :…, '1st Qu.:…, 'Median :…, 'Mean   :…, …, 'Median :…,
'Mean   :…, '3rd Qu.:…, 'Max.   :…]
```

### 0.3 行选择

[20]:
```
# 基于行所在位置筛选
df.iloc[[2,0,1]] # python 序数从 0 开始，2 代表第三行
df.loc[[2,0,1]] # 如果 index 未修改，效果与 iloc 的一致
```

[20]:

| | rownames | Hair | Eye | Sex | Freq |
|---|---|---|---|---|---|
| 2 | 3 | Red | Brown | Male | 10 |
| 0 | 1 | Black | Brown | Male | 32 |
| 1 | 2 | Brown | Brown | Male | 53 |

[21]:
```
# 单条件筛选，去掉.loc 效果一致
df.loc[df['Hair'] == 'Red']
```

[21]:

| | rownames | Hair | Eye | Sex | Freq |
|---|---|---|---|---|---|
| 2 | 3 | Red | Brown | Male | 10 |
| 6 | 7 | Red | Blue | Male | 10 |
| 10 | 11 | Red | Hazel | Male | 7 |
| 14 | 15 | Red | Green | Male | 7 |
| 18 | 19 | Red | Brown | Female | 16 |
| 22 | 23 | Red | Blue | Female | 7 |
| 26 | 27 | Red | Hazel | Female | 7 |
| 30 | 31 | Red | Green | Female | 7 |

[22]:
```
# pandas 多条件筛选时要用 |, &, ~ 分别代表 or, and, not; 且每个条件需要用括号区分
df.loc[(df['Hair'] == 'Black') &
       (df['Freq'] >= 10) &
       (df['Eye'].isin(['Brown', 'Blue']))]
```

[22]:

| | rownames | Hair | Eye | Sex | Freq |
|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 |
| 4 | 5 | Black | Blue | Male | 11 |
| 16 | 17 | Black | Brown | Female | 36 |

[23]:
```
%%R
# 基于行所在位置筛选，data.table 格式的 index 默认为 1 开始且
dt[c(3,1,2)]
```

```
   rownames  Hair   Eye  Sex Freq
1:        3   Red Brown Male   10
2:        1 Black Brown Male   32
3:        2 Brown Brown Male   53
```

[24]:
```
%%R
# 单条件筛选
dt[Hair == 'Red']
```

```
   rownames Hair   Eye    Sex Freq
1:        3  Red Brown   Male   10
```

```
2:        7  Red  Blue   Male   10
3:       11  Red Hazel   Male    7
4:       15  Red Green   Male    7
5:       19  Red Brown Female   16
6:       23  Red  Blue Female    7
7:       27  Red Hazel Female    7
8:       31  Red Green Female    7
```

[25]:
```
%%R
# 多条件筛选
dt[Hair == 'Black' &
   Freq >= 10 &
   Eye %in% c('Brown', 'Blue')]
```

```
   rownames  Hair   Eye    Sex Freq
1:        1 Black Brown   Male   32
2:        5 Black  Blue   Male   11
3:       17 Black Brown Female   36
```

### 0.4 行排序

[26]:
```
df.sort_values(['Sex', 'Freq'],
               ascending = [True, False] )
```

[26]:

| | rownames | Hair | Eye | Sex | Freq |
|---|---|---|---|---|---|
| 17 | 18 | Brown | Brown | Female | 66 |
| 23 | 24 | Blond | Blue | Female | 64 |
| 16 | 17 | Black | Brown | Female | 36 |
| 21 | 22 | Brown | Blue | Female | 34 |
| 25 | 26 | Brown | Hazel | Female | 29 |
| 18 | 19 | Red | Brown | Female | 16 |
| 29 | 30 | Brown | Green | Female | 14 |
| 20 | 21 | Black | Blue | Female | 9 |
| 31 | 32 | Blond | Green | Female | 8 |
| 22 | 23 | Red | Blue | Female | 7 |
| 26 | 27 | Red | Hazel | Female | 7 |
| 30 | 31 | Red | Green | Female | 7 |

```
24        25  Black  Hazel  Female     5
27        28  Blond  Hazel  Female     5
19        20  Blond  Brown  Female     4
28        29  Black  Green  Female     2
1          2  Brown  Brown    Male    53
5          6  Brown   Blue    Male    50
0          1  Black  Brown    Male    32
7          8  Blond   Blue    Male    30
9         10  Brown  Hazel    Male    25
13        14  Brown  Green    Male    15
4          5  Black   Blue    Male    11
2          3    Red  Brown    Male    10
6          7    Red   Blue    Male    10
8          9  Black  Hazel    Male    10
15        16  Blond  Green    Male     8
10        11    Red  Hazel    Male     7
14        15    Red  Green    Male     7
11        12  Blond  Hazel    Male     5
3          4  Blond  Brown    Male     3
12        13  Black  Green    Male     3
```

[28]:
```
%%R
dt[order(Sex, -Freq)]
```

```
    rownames  Hair   Eye    Sex Freq
 1:       18 Brown Brown Female   66
 2:       24 Blond  Blue Female   64
 3:       17 Black Brown Female   36
 4:       22 Brown  Blue Female   34
 5:       26 Brown Hazel Female   29
 6:       19   Red Brown Female   16
 7:       30 Brown Green Female   14
 8:       21 Black  Blue Female    9
 9:       32 Blond Green Female    8
10:       23   Red  Blue Female    7
11:       27   Red Hazel Female    7
12:       31   Red Green Female    7
```

```
13:       25 Black Hazel Female    5
14:       28 Blond Hazel Female    5
15:       20 Blond Brown Female    4
16:       29 Black Green Female    2
17:        2 Brown Brown   Male   53
18:        6 Brown  Blue   Male   50
19:        1 Black Brown   Male   32
20:        8 Blond  Blue   Male   30
21:       10 Brown Hazel   Male   25
22:       14 Brown Green   Male   15
23:        5 Black  Blue   Male   11
24:        3   Red Brown   Male   10
25:        7   Red  Blue   Male   10
26:        9 Black Hazel   Male   10
27:       16 Blond Green   Male    8
28:       11   Red Hazel   Male    7
29:       15   Red Green   Male    7
30:       12 Blond Hazel   Male    5
31:        4 Blond Brown   Male    3
32:       13 Black Green   Male    3
    rownames  Hair   Eye    Sex Freq
```

### 0.5 列选择

[29]:
```
df[['Hair', 'Freq']]
# or
df.loc[:, ['Eye', 'Sex']] # 选一列时也要保留 [], 否则与 df.Eye 一样为 series
```

[29]:
```
      Eye     Sex
0   Brown    Male
1   Brown    Male
2   Brown    Male
3   Brown    Male
4    Blue    Male
5    Blue    Male
6    Blue    Male
```

```
7    Blue    Male
8   Hazel    Male
9   Hazel    Male
10  Hazel    Male
11  Hazel    Male
12  Green    Male
13  Green    Male
14  Green    Male
15  Green    Male
16  Brown  Female
17  Brown  Female
18  Brown  Female
19  Brown  Female
20   Blue  Female
21   Blue  Female
22   Blue  Female
23   Blue  Female
24  Hazel  Female
25  Hazel  Female
26  Hazel  Female
27  Hazel  Female
28  Green  Female
29  Green  Female
30  Green  Female
31  Green  Female
```

[30]:
```
%%R
dt[, .(Hair, Freq)]
# or
dt[, c('Eye', 'Sex'), with=FALSE]
```

```
     Eye    Sex
1: Brown   Male
2: Brown   Male
3: Brown   Male
4: Brown   Male
5:  Blue   Male
```

```
 6:  Blue   Male
 7:  Blue   Male
 8:  Blue   Male
 9: Hazel   Male
10: Hazel   Male
11: Hazel   Male
12: Hazel   Male
13: Green   Male
14: Green   Male
15: Green   Male
16: Green   Male
17: Brown Female
18: Brown Female
19: Brown Female
20: Brown Female
21:  Blue Female
22:  Blue Female
23:  Blue Female
24:  Blue Female
25: Hazel Female
26: Hazel Female
27: Hazel Female
28: Hazel Female
29: Green Female
30: Green Female
31: Green Female
32: Green Female
      Eye    Sex
```

### 0.6 列新建

[32]:
```
# 新建一列
df = df.assign(nc = pd.Series(range(32)))
df.loc[:,'nc0'] = pd.Series(range(32), index=df.index)
df
```

[32]:

| | rownames | Hair | Eye | Sex | Freq | nc | nc0 |
|---|---|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 | 0 | 0 |
| 1 | 2 | Brown | Brown | Male | 53 | 1 | 1 |
| 2 | 3 | Red | Brown | Male | 10 | 2 | 2 |
| 3 | 4 | Blond | Brown | Male | 3 | 3 | 3 |
| 4 | 5 | Black | Blue | Male | 11 | 4 | 4 |
| 5 | 6 | Brown | Blue | Male | 50 | 5 | 5 |
| 6 | 7 | Red | Blue | Male | 10 | 6 | 6 |
| 7 | 8 | Blond | Blue | Male | 30 | 7 | 7 |
| 8 | 9 | Black | Hazel | Male | 10 | 8 | 8 |
| 9 | 10 | Brown | Hazel | Male | 25 | 9 | 9 |
| 10 | 11 | Red | Hazel | Male | 7 | 10 | 10 |
| 11 | 12 | Blond | Hazel | Male | 5 | 11 | 11 |
| 12 | 13 | Black | Green | Male | 3 | 12 | 12 |
| 13 | 14 | Brown | Green | Male | 15 | 13 | 13 |
| 14 | 15 | Red | Green | Male | 7 | 14 | 14 |
| 15 | 16 | Blond | Green | Male | 8 | 15 | 15 |
| 16 | 17 | Black | Brown | Female | 36 | 16 | 16 |
| 17 | 18 | Brown | Brown | Female | 66 | 17 | 17 |
| 18 | 19 | Red | Brown | Female | 16 | 18 | 18 |
| 19 | 20 | Blond | Brown | Female | 4 | 19 | 19 |
| 20 | 21 | Black | Blue | Female | 9 | 20 | 20 |
| 21 | 22 | Brown | Blue | Female | 34 | 21 | 21 |
| 22 | 23 | Red | Blue | Female | 7 | 22 | 22 |
| 23 | 24 | Blond | Blue | Female | 64 | 23 | 23 |
| 24 | 25 | Black | Hazel | Female | 5 | 24 | 24 |
| 25 | 26 | Brown | Hazel | Female | 29 | 25 | 25 |
| 26 | 27 | Red | Hazel | Female | 7 | 26 | 26 |
| 27 | 28 | Blond | Hazel | Female | 5 | 27 | 27 |
| 28 | 29 | Black | Green | Female | 2 | 28 | 28 |
| 29 | 30 | Brown | Green | Female | 14 | 29 | 29 |
| 30 | 31 | Red | Green | Female | 7 | 30 | 30 |
| 31 | 32 | Blond | Green | Female | 8 | 31 | 31 |

[34]:
```
# 新建多列
df = df.assign(
```

```
  nc1 = pd.Series(range(32)),
  nc2 = df.Hair + ',' + df.Eye
)
df
```

[34]:

| | rownames | Hair | Eye | Sex | Freq | nc | nc0 | nc1 | nc2 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 | 0 | 0 | 0 | Black,Brown |
| 1 | 2 | Brown | Brown | Male | 53 | 1 | 1 | 1 | Brown,Brown |
| 2 | 3 | Red | Brown | Male | 10 | 2 | 2 | 2 | Red,Brown |
| 3 | 4 | Blond | Brown | Male | 3 | 3 | 3 | 3 | Blond,Brown |
| 4 | 5 | Black | Blue | Male | 11 | 4 | 4 | 4 | Black,Blue |
| 5 | 6 | Brown | Blue | Male | 50 | 5 | 5 | 5 | Brown,Blue |
| 6 | 7 | Red | Blue | Male | 10 | 6 | 6 | 6 | Red,Blue |
| 7 | 8 | Blond | Blue | Male | 30 | 7 | 7 | 7 | Blond,Blue |
| 8 | 9 | Black | Hazel | Male | 10 | 8 | 8 | 8 | Black,Hazel |
| 9 | 10 | Brown | Hazel | Male | 25 | 9 | 9 | 9 | Brown,Hazel |
| 10 | 11 | Red | Hazel | Male | 7 | 10 | 10 | 10 | Red,Hazel |
| 11 | 12 | Blond | Hazel | Male | 5 | 11 | 11 | 11 | Blond,Hazel |
| 12 | 13 | Black | Green | Male | 3 | 12 | 12 | 12 | Black,Green |
| 13 | 14 | Brown | Green | Male | 15 | 13 | 13 | 13 | Brown,Green |
| 14 | 15 | Red | Green | Male | 7 | 14 | 14 | 14 | Red,Green |
| 15 | 16 | Blond | Green | Male | 8 | 15 | 15 | 15 | Blond,Green |
| 16 | 17 | Black | Brown | Female | 36 | 16 | 16 | 16 | Black,Brown |
| 17 | 18 | Brown | Brown | Female | 66 | 17 | 17 | 17 | Brown,Brown |
| 18 | 19 | Red | Brown | Female | 16 | 18 | 18 | 18 | Red,Brown |
| 19 | 20 | Blond | Brown | Female | 4 | 19 | 19 | 19 | Blond,Brown |
| 20 | 21 | Black | Blue | Female | 9 | 20 | 20 | 20 | Black,Blue |
| 21 | 22 | Brown | Blue | Female | 34 | 21 | 21 | 21 | Brown,Blue |
| 22 | 23 | Red | Blue | Female | 7 | 22 | 22 | 22 | Red,Blue |
| 23 | 24 | Blond | Blue | Female | 64 | 23 | 23 | 23 | Blond,Blue |
| 24 | 25 | Black | Hazel | Female | 5 | 24 | 24 | 24 | Black,Hazel |
| 25 | 26 | Brown | Hazel | Female | 29 | 25 | 25 | 25 | Brown,Hazel |
| 26 | 27 | Red | Hazel | Female | 7 | 26 | 26 | 26 | Red,Hazel |
| 27 | 28 | Blond | Hazel | Female | 5 | 27 | 27 | 27 | Blond,Hazel |
| 28 | 29 | Black | Green | Female | 2 | 28 | 28 | 28 | Black,Green |
| 29 | 30 | Brown | Green | Female | 14 | 29 | 29 | 29 | Brown,Green |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 30 | 31 | Red | Green | Female | 7 | 30 | 30 | 30 | Red,Green |
| 31 | 32 | Blond | Green | Female | 8 | 31 | 31 | 31 | Blond,Green |

[35]:
```
# 基于条件新建列
df = df.assign(nc3 = df.Freq.apply(lambda x: 1 if x >= 10 else 0))
df.loc[df.Freq >= 20, 'nc4'] = 2
df
```

[35]:

| | rownames | Hair | Eye | Sex | Freq | nc | nc0 | nc1 | nc2 | nc3 | nc4 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 | 0 | 0 | 0 | Black,Brown | 1 | 2.0 |
| 1 | 2 | Brown | Brown | Male | 53 | 1 | 1 | 1 | Brown,Brown | 1 | 2.0 |
| 2 | 3 | Red | Brown | Male | 10 | 2 | 2 | 2 | Red,Brown | 1 | NaN |
| 3 | 4 | Blond | Brown | Male | 3 | 3 | 3 | 3 | Blond,Brown | 0 | NaN |
| 4 | 5 | Black | Blue | Male | 11 | 4 | 4 | 4 | Black,Blue | 1 | NaN |
| 5 | 6 | Brown | Blue | Male | 50 | 5 | 5 | 5 | Brown,Blue | 1 | 2.0 |
| 6 | 7 | Red | Blue | Male | 10 | 6 | 6 | 6 | Red,Blue | 1 | NaN |
| 7 | 8 | Blond | Blue | Male | 30 | 7 | 7 | 7 | Blond,Blue | 1 | 2.0 |
| 8 | 9 | Black | Hazel | Male | 10 | 8 | 8 | 8 | Black,Hazel | 1 | NaN |
| 9 | 10 | Brown | Hazel | Male | 25 | 9 | 9 | 9 | Brown,Hazel | 1 | 2.0 |
| 10 | 11 | Red | Hazel | Male | 7 | 10 | 10 | 10 | Red,Hazel | 0 | NaN |
| 11 | 12 | Blond | Hazel | Male | 5 | 11 | 11 | 11 | Blond,Hazel | 0 | NaN |
| 12 | 13 | Black | Green | Male | 3 | 12 | 12 | 12 | Black,Green | 0 | NaN |
| 13 | 14 | Brown | Green | Male | 15 | 13 | 13 | 13 | Brown,Green | 1 | NaN |
| 14 | 15 | Red | Green | Male | 7 | 14 | 14 | 14 | Red,Green | 0 | NaN |
| 15 | 16 | Blond | Green | Male | 8 | 15 | 15 | 15 | Blond,Green | 0 | NaN |
| 16 | 17 | Black | Brown | Female | 36 | 16 | 16 | 16 | Black,Brown | 1 | 2.0 |
| 17 | 18 | Brown | Brown | Female | 66 | 17 | 17 | 17 | Brown,Brown | 1 | 2.0 |
| 18 | 19 | Red | Brown | Female | 16 | 18 | 18 | 18 | Red,Brown | 1 | NaN |
| 19 | 20 | Blond | Brown | Female | 4 | 19 | 19 | 19 | Blond,Brown | 0 | NaN |
| 20 | 21 | Black | Blue | Female | 9 | 20 | 20 | 20 | Black,Blue | 0 | NaN |
| 21 | 22 | Brown | Blue | Female | 34 | 21 | 21 | 21 | Brown,Blue | 1 | 2.0 |
| 22 | 23 | Red | Blue | Female | 7 | 22 | 22 | 22 | Red,Blue | 0 | NaN |
| 23 | 24 | Blond | Blue | Female | 64 | 23 | 23 | 23 | Blond,Blue | 1 | 2.0 |
| 24 | 25 | Black | Hazel | Female | 5 | 24 | 24 | 24 | Black,Hazel | 0 | NaN |
| 25 | 26 | Brown | Hazel | Female | 29 | 25 | 25 | 25 | Brown,Hazel | 1 | 2.0 |
| 26 | 27 | Red | Hazel | Female | 7 | 26 | 26 | 26 | Red,Hazel | 0 | NaN |
| 27 | 28 | Blond | Hazel | Female | 5 | 27 | 27 | 27 | Blond,Hazel | 0 | NaN |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 28 | 29 | Black | Green | Female | 2 | 28 | 28 | 28 | Black,Green | 0 | NaN |
| 29 | 30 | Brown | Green | Female | 14 | 29 | 29 | 29 | Brown,Green | 1 | NaN |
| 30 | 31 | Red | Green | Female | 7 | 30 | 30 | 30 | Red,Green | 0 | NaN |
| 31 | 32 | Blond | Green | Female | 8 | 31 | 31 | 31 | Blond,Green | 0 | NaN |

[36]:
```
# 基于函数新建多列
ncols = ['nc', 'nc0']
df.loc[:, ncols] = df[ncols].apply(lambda x: x**0.5+1)
df
```

[36]:

| | rownames | Hair | Eye | Sex | Freq | nc | nc0 | nc1 | \ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 | 1.000000 | 1.000000 | 0 | |
| 1 | 2 | Brown | Brown | Male | 53 | 2.000000 | 2.000000 | 1 | |
| 2 | 3 | Red | Brown | Male | 10 | 2.414214 | 2.414214 | 2 | |
| 3 | 4 | Blond | Brown | Male | 3 | 2.732051 | 2.732051 | 3 | |
| 4 | 5 | Black | Blue | Male | 11 | 3.000000 | 3.000000 | 4 | |
| 5 | 6 | Brown | Blue | Male | 50 | 3.236068 | 3.236068 | 5 | |
| 6 | 7 | Red | Blue | Male | 10 | 3.449490 | 3.449490 | 6 | |
| 7 | 8 | Blond | Blue | Male | 30 | 3.645751 | 3.645751 | 7 | |
| 8 | 9 | Black | Hazel | Male | 10 | 3.828427 | 3.828427 | 8 | |
| 9 | 10 | Brown | Hazel | Male | 25 | 4.000000 | 4.000000 | 9 | |
| 10 | 11 | Red | Hazel | Male | 7 | 4.162278 | 4.162278 | 10 | |
| 11 | 12 | Blond | Hazel | Male | 5 | 4.316625 | 4.316625 | 11 | |
| 12 | 13 | Black | Green | Male | 3 | 4.464102 | 4.464102 | 12 | |
| 13 | 14 | Brown | Green | Male | 15 | 4.605551 | 4.605551 | 13 | |
| 14 | 15 | Red | Green | Male | 7 | 4.741657 | 4.741657 | 14 | |
| 15 | 16 | Blond | Green | Male | 8 | 4.872983 | 4.872983 | 15 | |
| 16 | 17 | Black | Brown | Female | 36 | 5.000000 | 5.000000 | 16 | |
| 17 | 18 | Brown | Brown | Female | 66 | 5.123106 | 5.123106 | 17 | |
| 18 | 19 | Red | Brown | Female | 16 | 5.242641 | 5.242641 | 18 | |
| 19 | 20 | Blond | Brown | Female | 4 | 5.358899 | 5.358899 | 19 | |
| 20 | 21 | Black | Blue | Female | 9 | 5.472136 | 5.472136 | 20 | |
| 21 | 22 | Brown | Blue | Female | 34 | 5.582576 | 5.582576 | 21 | |
| 22 | 23 | Red | Blue | Female | 7 | 5.690416 | 5.690416 | 22 | |
| 23 | 24 | Blond | Blue | Female | 64 | 5.795832 | 5.795832 | 23 | |
| 24 | 25 | Black | Hazel | Female | 5 | 5.898979 | 5.898979 | 24 | |
| 25 | 26 | Brown | Hazel | Female | 29 | 6.000000 | 6.000000 | 25 | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 26 | 27 | Red | Hazel | Female | 7 | 6.099020 | 6.099020 | 26 |
| 27 | 28 | Blond | Hazel | Female | 5 | 6.196152 | 6.196152 | 27 |
| 28 | 29 | Black | Green | Female | 2 | 6.291503 | 6.291503 | 28 |
| 29 | 30 | Brown | Green | Female | 14 | 6.385165 | 6.385165 | 29 |
| 30 | 31 | Red | Green | Female | 7 | 6.477226 | 6.477226 | 30 |
| 31 | 32 | Blond | Green | Female | 8 | 6.567764 | 6.567764 | 31 |

| | nc2 | nc3 | nc4 |
|---|---|---|---|
| 0 | Black,Brown | 1 | 2.0 |
| 1 | Brown,Brown | 1 | 2.0 |
| 2 | Red,Brown | 1 | NaN |
| 3 | Blond,Brown | 0 | NaN |
| 4 | Black,Blue | 1 | NaN |
| 5 | Brown,Blue | 1 | 2.0 |
| 6 | Red,Blue | 1 | NaN |
| 7 | Blond,Blue | 1 | 2.0 |
| 8 | Black,Hazel | 1 | NaN |
| 9 | Brown,Hazel | 1 | 2.0 |
| 10 | Red,Hazel | 0 | NaN |
| 11 | Blond,Hazel | 0 | NaN |
| 12 | Black,Green | 0 | NaN |
| 13 | Brown,Green | 1 | NaN |
| 14 | Red,Green | 0 | NaN |
| 15 | Blond,Green | 0 | NaN |
| 16 | Black,Brown | 1 | 2.0 |
| 17 | Brown,Brown | 1 | 2.0 |
| 18 | Red,Brown | 1 | NaN |
| 19 | Blond,Brown | 0 | NaN |
| 20 | Black,Blue | 0 | NaN |
| 21 | Brown,Blue | 1 | 2.0 |
| 22 | Red,Blue | 0 | NaN |
| 23 | Blond,Blue | 1 | 2.0 |
| 24 | Black,Hazel | 0 | NaN |
| 25 | Brown,Hazel | 1 | 2.0 |
| 26 | Red,Hazel | 0 | NaN |
| 27 | Blond,Hazel | 0 | NaN |

```
28  Black,Green    0  NaN
29  Brown,Green    1  NaN
30    Red,Green    0  NaN
31  Blond,Green    0  NaN
```

[37]:
```
# 删除一列
df = df.drop('nc', axis=1)
df
```

[37]:

| | rownames | Hair | Eye | Sex | Freq | nc0 | nc1 | nc2 | nc3 | nc4 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 | 1.000000 | 0 | Black,Brown | 1 | 2.0 |
| 1 | 2 | Brown | Brown | Male | 53 | 2.000000 | 1 | Brown,Brown | 1 | 2.0 |
| 2 | 3 | Red | Brown | Male | 10 | 2.414214 | 2 | Red,Brown | 1 | NaN |
| 3 | 4 | Blond | Brown | Male | 3 | 2.732051 | 3 | Blond,Brown | 0 | NaN |
| 4 | 5 | Black | Blue | Male | 11 | 3.000000 | 4 | Black,Blue | 1 | NaN |
| 5 | 6 | Brown | Blue | Male | 50 | 3.236068 | 5 | Brown,Blue | 1 | 2.0 |
| 6 | 7 | Red | Blue | Male | 10 | 3.449490 | 6 | Red,Blue | 1 | NaN |
| 7 | 8 | Blond | Blue | Male | 30 | 3.645751 | 7 | Blond,Blue | 1 | 2.0 |
| 8 | 9 | Black | Hazel | Male | 10 | 3.828427 | 8 | Black,Hazel | 1 | NaN |
| 9 | 10 | Brown | Hazel | Male | 25 | 4.000000 | 9 | Brown,Hazel | 1 | 2.0 |
| 10 | 11 | Red | Hazel | Male | 7 | 4.162278 | 10 | Red,Hazel | 0 | NaN |
| 11 | 12 | Blond | Hazel | Male | 5 | 4.316625 | 11 | Blond,Hazel | 0 | NaN |
| 12 | 13 | Black | Green | Male | 3 | 4.464102 | 12 | Black,Green | 0 | NaN |
| 13 | 14 | Brown | Green | Male | 15 | 4.605551 | 13 | Brown,Green | 1 | NaN |
| 14 | 15 | Red | Green | Male | 7 | 4.741657 | 14 | Red,Green | 0 | NaN |
| 15 | 16 | Blond | Green | Male | 8 | 4.872983 | 15 | Blond,Green | 0 | NaN |
| 16 | 17 | Black | Brown | Female | 36 | 5.000000 | 16 | Black,Brown | 1 | 2.0 |
| 17 | 18 | Brown | Brown | Female | 66 | 5.123106 | 17 | Brown,Brown | 1 | 2.0 |
| 18 | 19 | Red | Brown | Female | 16 | 5.242641 | 18 | Red,Brown | 1 | NaN |
| 19 | 20 | Blond | Brown | Female | 4 | 5.358899 | 19 | Blond,Brown | 0 | NaN |
| 20 | 21 | Black | Blue | Female | 9 | 5.472136 | 20 | Black,Blue | 0 | NaN |
| 21 | 22 | Brown | Blue | Female | 34 | 5.582576 | 21 | Brown,Blue | 1 | 2.0 |
| 22 | 23 | Red | Blue | Female | 7 | 5.690416 | 22 | Red,Blue | 0 | NaN |
| 23 | 24 | Blond | Blue | Female | 64 | 5.795832 | 23 | Blond,Blue | 1 | 2.0 |
| 24 | 25 | Black | Hazel | Female | 5 | 5.898979 | 24 | Black,Hazel | 0 | NaN |
| 25 | 26 | Brown | Hazel | Female | 29 | 6.000000 | 25 | Brown,Hazel | 1 | 2.0 |
| 26 | 27 | Red | Hazel | Female | 7 | 6.099020 | 26 | Red,Hazel | 0 | NaN |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 27 | 28 | Blond | Hazel | Female | 5 | 6.196152 | 27 | Blond,Hazel | 0 | NaN |
| 28 | 29 | Black | Green | Female | 2 | 6.291503 | 28 | Black,Green | 0 | NaN |
| 29 | 30 | Brown | Green | Female | 14 | 6.385165 | 29 | Brown,Green | 1 | NaN |
| 30 | 31 | Red | Green | Female | 7 | 6.477226 | 30 | Red,Green | 0 | NaN |
| 31 | 32 | Blond | Green | Female | 8 | 6.567764 | 31 | Blond,Green | 0 | NaN |

[38]:
```
# 删除多列
df.drop(['nc0','nc1','nc2','nc3','nc4'], axis=1, inplace=True)
df
```

[38]:

| | rownames | Hair | Eye | Sex | Freq |
|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 |
| 1 | 2 | Brown | Brown | Male | 53 |
| 2 | 3 | Red | Brown | Male | 10 |
| 3 | 4 | Blond | Brown | Male | 3 |
| 4 | 5 | Black | Blue | Male | 11 |
| 5 | 6 | Brown | Blue | Male | 50 |
| 6 | 7 | Red | Blue | Male | 10 |
| 7 | 8 | Blond | Blue | Male | 30 |
| 8 | 9 | Black | Hazel | Male | 10 |
| 9 | 10 | Brown | Hazel | Male | 25 |
| 10 | 11 | Red | Hazel | Male | 7 |
| 11 | 12 | Blond | Hazel | Male | 5 |
| 12 | 13 | Black | Green | Male | 3 |
| 13 | 14 | Brown | Green | Male | 15 |
| 14 | 15 | Red | Green | Male | 7 |
| 15 | 16 | Blond | Green | Male | 8 |
| 16 | 17 | Black | Brown | Female | 36 |
| 17 | 18 | Brown | Brown | Female | 66 |
| 18 | 19 | Red | Brown | Female | 16 |
| 19 | 20 | Blond | Brown | Female | 4 |
| 20 | 21 | Black | Blue | Female | 9 |
| 21 | 22 | Brown | Blue | Female | 34 |
| 22 | 23 | Red | Blue | Female | 7 |
| 23 | 24 | Blond | Blue | Female | 64 |
| 24 | 25 | Black | Hazel | Female | 5 |
| 25 | 26 | Brown | Hazel | Female | 29 |

```
26        27    Red  Hazel  Female     7
27        28  Blond  Hazel  Female     5
28        29  Black  Green  Female     2
29        30  Brown  Green  Female    14
30        31    Red  Green  Female     7
31        32  Blond  Green  Female     8
```

[39]:
```
%%R
# 新建一列
dt[, nc := .I] # .I .N .SD 为特殊符号，查看帮助?`.I`
# .SD 是指数据中的子集，具体功能是对列进行筛选，可以配合 by 一起使用。
# .SDcols 可以选择列的子集。
# .N 类似 nrow() 函数，即返回每组的长度，也就是最大行号。
# .I 类似 seq_len(nrow(x))，就是返回行号。
# .GRP 生成分组序号，在根据多变量分组的时候很有用。
dt[,'nc0'] = 1:32
dt
```

```
    rownames  Hair   Eye    Sex Freq nc nc0
 1:        1 Black Brown   Male   32  1   1
 2:        2 Brown Brown   Male   53  2   2
 3:        3   Red Brown   Male   10  3   3
 4:        4 Blond Brown   Male    3  4   4
 5:        5 Black  Blue   Male   11  5   5
 6:        6 Brown  Blue   Male   50  6   6
 7:        7   Red  Blue   Male   10  7   7
 8:        8 Blond  Blue   Male   30  8   8
 9:        9 Black Hazel   Male   10  9   9
10:       10 Brown Hazel   Male   25 10  10
11:       11   Red Hazel   Male    7 11  11
12:       12 Blond Hazel   Male    5 12  12
13:       13 Black Green   Male    3 13  13
14:       14 Brown Green   Male   15 14  14
15:       15   Red Green   Male    7 15  15
16:       16 Blond Green   Male    8 16  16
17:       17 Black Brown Female   36 17  17
18:       18 Brown Brown Female   66 18  18
```

```
19:       19   Red Brown Female   16 19  19
20:       20 Blond Brown Female    4 20  20
21:       21 Black  Blue Female    9 21  21
22:       22 Brown  Blue Female   34 22  22
23:       23   Red  Blue Female    7 23  23
24:       24 Blond  Blue Female   64 24  24
25:       25 Black Hazel Female    5 25  25
26:       26 Brown Hazel Female   29 26  26
27:       27   Red Hazel Female    7 27  27
28:       28 Blond Hazel Female    5 28  28
29:       29 Black Green Female    2 29  29
30:       30 Brown Green Female   14 30  30
31:       31   Red Green Female    7 31  31
32:       32 Blond Green Female    8 32  32
    rownames  Hair   Eye    Sex Freq nc nc0
```

[40]:
```
%%R
# 新建多列
dt[, `:=`(
  nc1 = 1:32,
  nc2 = paste(Hair, Eye, sep=',')
)]
dt
```

```
    rownames  Hair   Eye    Sex Freq nc nc0 nc1         nc2
 1:        1 Black Brown   Male   32  1   1   1 Black,Brown
 2:        2 Brown Brown   Male   53  2   2   2 Brown,Brown
 3:        3   Red Brown   Male   10  3   3   3   Red,Brown
 4:        4 Blond Brown   Male    3  4   4   4 Blond,Brown
 5:        5 Black  Blue   Male   11  5   5   5  Black,Blue
 6:        6 Brown  Blue   Male   50  6   6   6  Brown,Blue
 7:        7   Red  Blue   Male   10  7   7   7    Red,Blue
 8:        8 Blond  Blue   Male   30  8   8   8  Blond,Blue
 9:        9 Black Hazel   Male   10  9   9   9 Black,Hazel
10:       10 Brown Hazel   Male   25 10  10  10 Brown,Hazel
11:       11   Red Hazel   Male    7 11  11  11   Red,Hazel
12:       12 Blond Hazel   Male    5 12  12  12 Blond,Hazel
```

```
13:       13 Black Green   Male    3 13  13  13 Black,Green
14:       14 Brown Green   Male   15 14  14  14 Brown,Green
15:       15   Red Green   Male    7 15  15  15   Red,Green
16:       16 Blond Green   Male    8 16  16  16 Blond,Green
17:       17 Black Brown Female   36 17  17  17 Black,Brown
18:       18 Brown Brown Female   66 18  18  18 Brown,Brown
19:       19   Red Brown Female   16 19  19  19   Red,Brown
20:       20 Blond Brown Female    4 20  20  20 Blond,Brown
21:       21 Black  Blue Female    9 21  21  21  Black,Blue
22:       22 Brown  Blue Female   34 22  22  22  Brown,Blue
23:       23   Red  Blue Female    7 23  23  23    Red,Blue
24:       24 Blond  Blue Female   64 24  24  24  Blond,Blue
25:       25 Black Hazel Female    5 25  25  25 Black,Hazel
26:       26 Brown Hazel Female   29 26  26  26 Brown,Hazel
27:       27   Red Hazel Female    7 27  27  27   Red,Hazel
28:       28 Blond Hazel Female    5 28  28  28 Blond,Hazel
29:       29 Black Green Female    2 29  29  29 Black,Green
30:       30 Brown Green Female   14 30  30  30 Brown,Green
31:       31   Red Green Female    7 31  31  31   Red,Green
32:       32 Blond Green Female    8 32  32  32 Blond,Green
    rownames  Hair   Eye    Sex Freq nc nc0 nc1         nc2
```

[41]:
```
%%R
# 基于条件新建列
dt[, nc3 := ifelse(Freq >= 10, 1, 0)]
dt[Freq >= 20, nc4 := 2]
dt
```

```
   rownames  Hair   Eye    Sex Freq nc nc0 nc1         nc2 nc3 nc4
1:        1 Black Brown   Male   32  1   1   1 Black,Brown   1   2
2:        2 Brown Brown   Male   53  2   2   2 Brown,Brown   1   2
3:        3   Red Brown   Male   10  3   3   3   Red,Brown   1  NA
4:        4 Blond Brown   Male    3  4   4   4 Blond,Brown   0  NA
5:        5 Black  Blue   Male   11  5   5   5  Black,Blue   1  NA
6:        6 Brown  Blue   Male   50  6   6   6  Brown,Blue   1   2
7:        7   Red  Blue   Male   10  7   7   7    Red,Blue   1  NA
8:        8 Blond  Blue   Male   30  8   8   8  Blond,Blue   1   2
```

```
 9:        9 Black Hazel   Male   10  9   9   9 Black,Hazel   1  NA
10:       10 Brown Hazel   Male   25 10  10  10 Brown,Hazel   1   2
11:       11   Red Hazel   Male    7 11  11  11   Red,Hazel   0  NA
12:       12 Blond Hazel   Male    5 12  12  12 Blond,Hazel   0  NA
13:       13 Black Green   Male    3 13  13  13 Black,Green   0  NA
14:       14 Brown Green   Male   15 14  14  14 Brown,Green   1  NA
15:       15   Red Green   Male    7 15  15  15   Red,Green   0  NA
16:       16 Blond Green   Male    8 16  16  16 Blond,Green   0  NA
17:       17 Black Brown Female   36 17  17  17 Black,Brown   1   2
18:       18 Brown Brown Female   66 18  18  18 Brown,Brown   1   2
19:       19   Red Brown Female   16 19  19  19   Red,Brown   1  NA
20:       20 Blond Brown Female    4 20  20  20 Blond,Brown   0  NA
21:       21 Black  Blue Female    9 21  21  21  Black,Blue   0  NA
22:       22 Brown  Blue Female   34 22  22  22  Brown,Blue   1   2
23:       23   Red  Blue Female    7 23  23  23    Red,Blue   0  NA
24:       24 Blond  Blue Female   64 24  24  24  Blond,Blue   1   2
25:       25 Black Hazel Female    5 25  25  25 Black,Hazel   0  NA
26:       26 Brown Hazel Female   29 26  26  26 Brown,Hazel   1   2
27:       27   Red Hazel Female    7 27  27  27   Red,Hazel   0  NA
28:       28 Blond Hazel Female    5 28  28  28 Blond,Hazel   0  NA
29:       29 Black Green Female    2 29  29  29 Black,Green   0  NA
30:       30 Brown Green Female   14 30  30  30 Brown,Green   1  NA
31:       31   Red Green Female    7 31  31  31   Red,Green   0  NA
32:       32 Blond Green Female    8 32  32  32 Blond,Green   0  NA
    rownames  Hair   Eye    Sex Freq nc nc0 nc1         nc2 nc3 nc4
```

[42]:
```
%%R
# 基于函数新建多列
ncols = c('nc', 'nc0')
dt[,
   (ncols) := lapply(.SD, function(x) x^0.5+1),
   .SDcols = ncols]
dt
```

```
    rownames  Hair   Eye    Sex Freq       nc      nc0 nc1         nc2 nc3 nc4
 1:        1 Black Brown   Male   32 2.000000 2.000000   1 Black,Brown   1   2
 2:        2 Brown Brown   Male   53 2.414214 2.414214   2 Brown,Brown   1   2
```

```
 3:        3   Red Brown   Male   10 2.732051 2.732051   3   Red,Brown   1  NA
 4:        4 Blond Brown   Male    3 3.000000 3.000000   4 Blond,Brown   0  NA
 5:        5 Black  Blue   Male   11 3.236068 3.236068   5  Black,Blue   1  NA
 6:        6 Brown  Blue   Male   50 3.449490 3.449490   6  Brown,Blue   1   2
 7:        7   Red  Blue   Male   10 3.645751 3.645751   7    Red,Blue   1  NA
 8:        8 Blond  Blue   Male   30 3.828427 3.828427   8  Blond,Blue   1   2
 9:        9 Black Hazel   Male   10 4.000000 4.000000   9 Black,Hazel   1  NA
10:       10 Brown Hazel   Male   25 4.162278 4.162278  10 Brown,Hazel   1   2
11:       11   Red Hazel   Male    7 4.316625 4.316625  11   Red,Hazel   0  NA
12:       12 Blond Hazel   Male    5 4.464102 4.464102  12 Blond,Hazel   0  NA
13:       13 Black Green   Male    3 4.605551 4.605551  13 Black,Green   0  NA
14:       14 Brown Green   Male   15 4.741657 4.741657  14 Brown,Green   1  NA
15:       15   Red Green   Male    7 4.872983 4.872983  15   Red,Green   0  NA
16:       16 Blond Green   Male    8 5.000000 5.000000  16 Blond,Green   0  NA
17:       17 Black Brown Female   36 5.123106 5.123106  17 Black,Brown   1   2
18:       18 Brown Brown Female   66 5.242641 5.242641  18 Brown,Brown   1   2
19:       19   Red Brown Female   16 5.358899 5.358899  19   Red,Brown   1  NA
20:       20 Blond Brown Female    4 5.472136 5.472136  20 Blond,Brown   0  NA
21:       21 Black  Blue Female    9 5.582576 5.582576  21  Black,Blue   0  NA
22:       22 Brown  Blue Female   34 5.690416 5.690416  22  Brown,Blue   1   2
23:       23   Red  Blue Female    7 5.795832 5.795832  23    Red,Blue   0  NA
24:       24 Blond  Blue Female   64 5.898979 5.898979  24  Blond,Blue   1   2
25:       25 Black Hazel Female    5 6.000000 6.000000  25 Black,Hazel   0  NA
26:       26 Brown Hazel Female   29 6.099020 6.099020  26 Brown,Hazel   1   2
27:       27   Red Hazel Female    7 6.196152 6.196152  27   Red,Hazel   0  NA
28:       28 Blond Hazel Female    5 6.291503 6.291503  28 Blond,Hazel   0  NA
29:       29 Black Green Female    2 6.385165 6.385165  29 Black,Green   0  NA
30:       30 Brown Green Female   14 6.477226 6.477226  30 Brown,Green   1  NA
31:       31   Red Green Female    7 6.567764 6.567764  31   Red,Green   0  NA
32:       32 Blond Green Female    8 6.656854 6.656854  32 Blond,Green   0  NA
    rownames  Hair   Eye    Sex Freq       nc      nc0 nc1         nc2 nc3 nc4
```

[43]:
```
%%R
# 删除一列
dt[, nc := NULL]
```

```
    rownames  Hair   Eye    Sex Freq      nc0 nc1         nc2 nc3 nc4
```

```
 1:        1 Black Brown   Male   32 2.000000   1 Black,Brown   1   2
 2:        2 Brown Brown   Male   53 2.414214   2 Brown,Brown   1   2
 3:        3   Red Brown   Male   10 2.732051   3   Red,Brown   1  NA
 4:        4 Blond Brown   Male    3 3.000000   4 Blond,Brown   0  NA
 5:        5 Black  Blue   Male   11 3.236068   5  Black,Blue   1  NA
 6:        6 Brown  Blue   Male   50 3.449490   6  Brown,Blue   1   2
 7:        7   Red  Blue   Male   10 3.645751   7    Red,Blue   1  NA
 8:        8 Blond  Blue   Male   30 3.828427   8  Blond,Blue   1   2
 9:        9 Black Hazel   Male   10 4.000000   9 Black,Hazel   1  NA
10:       10 Brown Hazel   Male   25 4.162278  10 Brown,Hazel   1   2
11:       11   Red Hazel   Male    7 4.316625  11   Red,Hazel   0  NA
12:       12 Blond Hazel   Male    5 4.464102  12 Blond,Hazel   0  NA
13:       13 Black Green   Male    3 4.605551  13 Black,Green   0  NA
14:       14 Brown Green   Male   15 4.741657  14 Brown,Green   1  NA
15:       15   Red Green   Male    7 4.872983  15   Red,Green   0  NA
16:       16 Blond Green   Male    8 5.000000  16 Blond,Green   0  NA
17:       17 Black Brown Female   36 5.123106  17 Black,Brown   1   2
18:       18 Brown Brown Female   66 5.242641  18 Brown,Brown   1   2
19:       19   Red Brown Female   16 5.358899  19   Red,Brown   1  NA
20:       20 Blond Brown Female    4 5.472136  20 Blond,Brown   0  NA
21:       21 Black  Blue Female    9 5.582576  21  Black,Blue   0  NA
22:       22 Brown  Blue Female   34 5.690416  22  Brown,Blue   1   2
23:       23   Red  Blue Female    7 5.795832  23    Red,Blue   0  NA
24:       24 Blond  Blue Female   64 5.898979  24  Blond,Blue   1   2
25:       25 Black Hazel Female    5 6.000000  25 Black,Hazel   0  NA
26:       26 Brown Hazel Female   29 6.099020  26 Brown,Hazel   1   2
27:       27   Red Hazel Female    7 6.196152  27   Red,Hazel   0  NA
28:       28 Blond Hazel Female    5 6.291503  28 Blond,Hazel   0  NA
29:       29 Black Green Female    2 6.385165  29 Black,Green   0  NA
30:       30 Brown Green Female   14 6.477226  30 Brown,Green   1  NA
31:       31   Red Green Female    7 6.567764  31   Red,Green   0  NA
32:       32 Blond Green Female    8 6.656854  32 Blond,Green   0  NA
    rownames  Hair   Eye    Sex Freq      nc0 nc1         nc2 nc3 nc4
```

[44]:
```
%%R
# 删除多列
dt[, (c('nc0','nc1','nc2','nc3','nc4')) := NULL]
```

```
    rownames  Hair   Eye    Sex Freq
 1:        1 Black Brown   Male   32
 2:        2 Brown Brown   Male   53
 3:        3   Red Brown   Male   10
 4:        4 Blond Brown   Male    3
 5:        5 Black  Blue   Male   11
 6:        6 Brown  Blue   Male   50
 7:        7   Red  Blue   Male   10
 8:        8 Blond  Blue   Male   30
 9:        9 Black Hazel   Male   10
10:       10 Brown Hazel   Male   25
11:       11   Red Hazel   Male    7
12:       12 Blond Hazel   Male    5
13:       13 Black Green   Male    3
14:       14 Brown Green   Male   15
15:       15   Red Green   Male    7
16:       16 Blond Green   Male    8
17:       17 Black Brown Female   36
18:       18 Brown Brown Female   66
19:       19   Red Brown Female   16
20:       20 Blond Brown Female    4
21:       21 Black  Blue Female    9
22:       22 Brown  Blue Female   34
23:       23   Red  Blue Female    7
24:       24 Blond  Blue Female   64
25:       25 Black Hazel Female    5
26:       26 Brown Hazel Female   29
27:       27   Red Hazel Female    7
28:       28 Blond Hazel Female    5
29:       29 Black Green Female    2
30:       30 Brown Green Female   14
31:       31   Red Green Female    7
32:       32 Blond Green Female    8
    rownames  Hair   Eye    Sex Freq
```

## 0.7 列计算

[45]:
```
# 对一列进行计算
df.Freq.max() # 最大值
df.Eye.unique() # 唯一值
df.Eye.value_counts() # 计数
```

[45]:
```
Brown    8
Blue     8
Hazel    8
Green    8
Name: Eye, dtype: int64
```

[46]:
```
# 对多列进行计算
## 所有列的最大值
df.max()
```

[46]:
```
rownames       32
Hair          Red
Eye         Hazel
Sex          Male
Freq           66
dtype: object
```

[47]:
```
## 所有列的缺失率
df.isnull().mean()
```

[47]:
```
rownames    0.0
Hair        0.0
Eye         0.0
Sex         0.0
Freq        0.0
dtype: float64
```

[48]:
```
## 对部分列计算缺失率，且可扩展到其他函数
sel_cols = ['Hair', 'Sex', 'Freq']
df[sel_cols].apply(lambda x: x.isnull().mean())
```

[48]:
```
Hair    0.0
Sex     0.0
Freq    0.0
dtype: float64
```

[49]:
```
%%R
# 对一列进行计算
dt[, max(Freq)] # 最大值
dt[, unique(Eye)] # 唯一值
dt[, table(Eye)] # 计数
```

```
Eye
 Blue Brown Green Hazel
    8     8     8     8
```

[50]:
```
%%R
# 对多列进行计算
## 所有列的最大值
dt[, lapply(.SD, max)]
```

```
   rownames Hair   Eye  Sex Freq
1:       32  Red Hazel Male   66
```

[51]:
```
%%R
## 所有列的缺失率
dt[, lapply(.SD, function(x) mean(is.na(x)))]
```

```
   rownames Hair Eye Sex Freq
1:        0    0   0   0    0
```

[52]:
```
%%R
## 对部分列计算缺失率，且可扩展到其他函数
sel_cols = c('Hair', 'Sex', 'Freq')
dt[, lapply(.SD, function(x) mean(is.na(x))), .SDcols = sel_cols]
```

```
   Hair Sex Freq
1:    0   0    0
```

## 0.8 分组数据操作

[53]:
```
# 分组行操作
## 行选择
df.groupby('Sex').head(1) # 每组的第一行
df.groupby('Sex').tail(1) # 每组的最后一行
```

[53]:

| | rownames | Hair | Eye | Sex | Freq |
|---|---|---|---|---|---|
| 15 | 16 | Blond | Green | Male | 8 |
| 31 | 32 | Blond | Green | Female | 8 |

[55]:
```
# 分组列操作
## 分组列新建
df.loc[:,'freq_total'] = df.groupby('Sex')['Freq'].transform(sum)
df
```

[55]:

| | rownames | Hair | Eye | Sex | Freq | freq_total |
|---|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 | 279 |
| 1 | 2 | Brown | Brown | Male | 53 | 279 |
| 2 | 3 | Red | Brown | Male | 10 | 279 |
| 3 | 4 | Blond | Brown | Male | 3 | 279 |
| 4 | 5 | Black | Blue | Male | 11 | 279 |
| 5 | 6 | Brown | Blue | Male | 50 | 279 |
| 6 | 7 | Red | Blue | Male | 10 | 279 |
| 7 | 8 | Blond | Blue | Male | 30 | 279 |
| 8 | 9 | Black | Hazel | Male | 10 | 279 |
| 9 | 10 | Brown | Hazel | Male | 25 | 279 |
| 10 | 11 | Red | Hazel | Male | 7 | 279 |
| 11 | 12 | Blond | Hazel | Male | 5 | 279 |
| 12 | 13 | Black | Green | Male | 3 | 279 |
| 13 | 14 | Brown | Green | Male | 15 | 279 |
| 14 | 15 | Red | Green | Male | 7 | 279 |
| 15 | 16 | Blond | Green | Male | 8 | 279 |
| 16 | 17 | Black | Brown | Female | 36 | 313 |
| 17 | 18 | Brown | Brown | Female | 66 | 313 |
| 18 | 19 | Red | Brown | Female | 16 | 313 |
| 19 | 20 | Blond | Brown | Female | 4 | 313 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 20 | 21 | Black | Blue | Female | 9 | 313 |
| 21 | 22 | Brown | Blue | Female | 34 | 313 |
| 22 | 23 | Red | Blue | Female | 7 | 313 |
| 23 | 24 | Blond | Blue | Female | 64 | 313 |
| 24 | 25 | Black | Hazel | Female | 5 | 313 |
| 25 | 26 | Brown | Hazel | Female | 29 | 313 |
| 26 | 27 | Red | Hazel | Female | 7 | 313 |
| 27 | 28 | Blond | Hazel | Female | 5 | 313 |
| 28 | 29 | Black | Green | Female | 2 | 313 |
| 29 | 30 | Brown | Green | Female | 14 | 313 |
| 30 | 31 | Red | Green | Female | 7 | 313 |
| 31 | 32 | Blond | Green | Female | 8 | 313 |

[59]:
```
## 分组列计算
df.groupby('Sex').agg({'Freq':'sum'}).rename(columns={'Freq':'freq_total'}).
 ↪reset_index()
df
```

[59]:

| | rownames | Hair | Eye | Sex | Freq | freq_total |
|---|---|---|---|---|---|---|
| 0 | 1 | Black | Brown | Male | 32 | 279 |
| 1 | 2 | Brown | Brown | Male | 53 | 279 |
| 2 | 3 | Red | Brown | Male | 10 | 279 |
| 3 | 4 | Blond | Brown | Male | 3 | 279 |
| 4 | 5 | Black | Blue | Male | 11 | 279 |
| 5 | 6 | Brown | Blue | Male | 50 | 279 |
| 6 | 7 | Red | Blue | Male | 10 | 279 |
| 7 | 8 | Blond | Blue | Male | 30 | 279 |
| 8 | 9 | Black | Hazel | Male | 10 | 279 |
| 9 | 10 | Brown | Hazel | Male | 25 | 279 |
| 10 | 11 | Red | Hazel | Male | 7 | 279 |
| 11 | 12 | Blond | Hazel | Male | 5 | 279 |
| 12 | 13 | Black | Green | Male | 3 | 279 |
| 13 | 14 | Brown | Green | Male | 15 | 279 |
| 14 | 15 | Red | Green | Male | 7 | 279 |
| 15 | 16 | Blond | Green | Male | 8 | 279 |
| 16 | 17 | Black | Brown | Female | 36 | 313 |

```
17        18  Brown  Brown  Female    66         313
18        19    Red  Brown  Female    16         313
19        20  Blond  Brown  Female     4         313
20        21  Black   Blue  Female     9         313
21        22  Brown   Blue  Female    34         313
22        23    Red   Blue  Female     7         313
23        24  Blond   Blue  Female    64         313
24        25  Black  Hazel  Female     5         313
25        26  Brown  Hazel  Female    29         313
26        27    Red  Hazel  Female     7         313
27        28  Blond  Hazel  Female     5         313
28        29  Black  Green  Female     2         313
29        30  Brown  Green  Female    14         313
30        31    Red  Green  Female     7         313
31        32  Blond  Green  Female     8         313
```

[60]:
```
%%R
# 分组行操作
## 行选择
dt[, .SD[1], by = 'Sex'] # 每组的第一行
dt[, .SD[.N], by = 'Sex'] # 每组的最后一行
```

```
     Sex rownames  Hair   Eye Freq
1:   Male       16 Blond Green    8
2: Female       32 Blond Green    8
```

[61]:
```
%%R
# 分组列操作
## 分组列新建
dt[, freq_total := sum(Freq), by = 'Sex']
```

```
   rownames  Hair   Eye    Sex Freq freq_total
1:        1 Black Brown   Male   32        279
2:        2 Brown Brown   Male   53        279
3:        3   Red Brown   Male   10        279
4:        4 Blond Brown   Male    3        279
5:        5 Black  Blue   Male   11        279
```

```
 6:        6 Brown  Blue   Male   50        279
 7:        7   Red  Blue   Male   10        279
 8:        8 Blond  Blue   Male   30        279
 9:        9 Black Hazel   Male   10        279
10:       10 Brown Hazel   Male   25        279
11:       11   Red Hazel   Male    7        279
12:       12 Blond Hazel   Male    5        279
13:       13 Black Green   Male    3        279
14:       14 Brown Green   Male   15        279
15:       15   Red Green   Male    7        279
16:       16 Blond Green   Male    8        279
17:       17 Black Brown Female   36        313
18:       18 Brown Brown Female   66        313
19:       19   Red Brown Female   16        313
20:       20 Blond Brown Female    4        313
21:       21 Black  Blue Female    9        313
22:       22 Brown  Blue Female   34        313
23:       23   Red  Blue Female    7        313
24:       24 Blond  Blue Female   64        313
25:       25 Black Hazel Female    5        313
26:       26 Brown Hazel Female   29        313
27:       27   Red Hazel Female    7        313
28:       28 Blond Hazel Female    5        313
29:       29 Black Green Female    2        313
30:       30 Brown Green Female   14        313
31:       31   Red Green Female    7        313
32:       32 Blond Green Female    8        313
    rownames  Hair   Eye    Sex Freq freq_total
```

[62]:
```
%%R
## 分组列计算
dt[, .(freq_total = sum(Freq)), by = 'Sex'][]
```

```
      Sex freq_total
1:   Male        279
2: Female        313
```

## 0.9 长宽表转换

[63]:
```
# 长表转宽表
df_w = pd.pivot_table(df, index=['Hair','Sex'], columns='Eye', values='Freq',␣
 ↪aggfunc = sum).reset_index()
df_w
```

[63]:

| Eye | Hair | Sex | Blue | Brown | Green | Hazel |
|---|---|---|---|---|---|---|
| 0 | Black | Female | 9 | 36 | 2 | 5 |
| 1 | Black | Male | 11 | 32 | 3 | 10 |
| 2 | Blond | Female | 64 | 4 | 8 | 5 |
| 3 | Blond | Male | 30 | 3 | 8 | 5 |
| 4 | Brown | Female | 34 | 66 | 14 | 29 |
| 5 | Brown | Male | 50 | 53 | 15 | 25 |
| 6 | Red | Female | 7 | 16 | 7 | 7 |
| 7 | Red | Male | 10 | 10 | 7 | 7 |

[64]:
```
# 宽表转长表
df_l = pd.melt(df_w, id_vars = ['Hair','Sex'], var_name='Freq')
df_l
```

[64]:

| | Hair | Sex | Freq | value |
|---|---|---|---|---|
| 0 | Black | Female | Blue | 9 |
| 1 | Black | Male | Blue | 11 |
| 2 | Blond | Female | Blue | 64 |
| 3 | Blond | Male | Blue | 30 |
| 4 | Brown | Female | Blue | 34 |
| 5 | Brown | Male | Blue | 50 |
| 6 | Red | Female | Blue | 7 |
| 7 | Red | Male | Blue | 10 |
| 8 | Black | Female | Brown | 36 |
| 9 | Black | Male | Brown | 32 |
| 10 | Blond | Female | Brown | 4 |
| 11 | Blond | Male | Brown | 3 |
| 12 | Brown | Female | Brown | 66 |
| 13 | Brown | Male | Brown | 53 |
| 14 | Red | Female | Brown | 16 |

```
15    Red    Male  Brown     10
16  Black  Female  Green      2
17  Black    Male  Green      3
18  Blond  Female  Green      8
19  Blond    Male  Green      8
20  Brown  Female  Green     14
21  Brown    Male  Green     15
22    Red  Female  Green      7
23    Red    Male  Green      7
24  Black  Female  Hazel      5
25  Black    Male  Hazel     10
26  Blond  Female  Hazel      5
27  Blond    Male  Hazel      5
28  Brown  Female  Hazel     29
29  Brown    Male  Hazel     25
30    Red  Female  Hazel      7
31    Red    Male  Hazel      7
```

[66]:
```
%%R
# 长表转宽表
dt_w = dcast(dt, Hair+Sex~Eye, value.var = 'Freq', fun.aggregate = sum)
dt_w
```

```
    Hair    Sex Blue Brown Green Hazel
1: Black Female    9    36     2     5
2: Black   Male   11    32     3    10
3: Blond Female   64     4     8     5
4: Blond   Male   30     3     8     5
5: Brown Female   34    66    14    29
6: Brown   Male   50    53    15    25
7:   Red Female    7    16     7     7
8:   Red   Male   10    10     7     7
```

[67]:
```
%%R
# 宽表转长表
dt_l = melt(dt_w, id = c('Hair','Sex'), variable.name = 'Eye', value.name =␣
 ↪'Freq')
```

```
dt_l
```

```
     Hair    Sex   Eye Freq
 1: Black Female  Blue    9
 2: Black   Male  Blue   11
 3: Blond Female  Blue   64
 4: Blond   Male  Blue   30
 5: Brown Female  Blue   34
 6: Brown   Male  Blue   50
 7:   Red Female  Blue    7
 8:   Red   Male  Blue   10
 9: Black Female Brown   36
10: Black   Male Brown   32
11: Blond Female Brown    4
12: Blond   Male Brown    3
13: Brown Female Brown   66
14: Brown   Male Brown   53
15:   Red Female Brown   16
16:   Red   Male Brown   10
17: Black Female Green    2
18: Black   Male Green    3
19: Blond Female Green    8
20: Blond   Male Green    8
21: Brown Female Green   14
22: Brown   Male Green   15
23:   Red Female Green    7
24:   Red   Male Green    7
25: Black Female Hazel    5
26: Black   Male Hazel   10
27: Blond Female Hazel    5
28: Blond   Male Hazel    5
29: Brown Female Hazel   29
30: Brown   Male Hazel   25
31:   Red Female Hazel    7
32:   Red   Male Hazel    7
     Hair    Sex   Eye Freq
```

## 0.10 行列切割合并

[70]:
```
# 一行切割为多行
dfr = df.groupby(['Hair','Sex'])['Eye'].apply(lambda x: ','.join(x)).
 ↪reset_index()
# dfr
dfr.assign(Eye = dfr['Eye'].str.split(',')).explode('Eye')
```

[70]:

| | Hair | Sex | Eye |
|---|---|---|---|
| 0 | Black | Female | Brown |
| 0 | Black | Female | Blue |
| 0 | Black | Female | Hazel |
| 0 | Black | Female | Green |
| 1 | Black | Male | Brown |
| 1 | Black | Male | Blue |
| 1 | Black | Male | Hazel |
| 1 | Black | Male | Green |
| 2 | Blond | Female | Brown |
| 2 | Blond | Female | Blue |
| 2 | Blond | Female | Hazel |
| 2 | Blond | Female | Green |
| 3 | Blond | Male | Brown |
| 3 | Blond | Male | Blue |
| 3 | Blond | Male | Hazel |
| 3 | Blond | Male | Green |
| 4 | Brown | Female | Brown |
| 4 | Brown | Female | Blue |
| 4 | Brown | Female | Hazel |
| 4 | Brown | Female | Green |
| 5 | Brown | Male | Brown |
| 5 | Brown | Male | Blue |
| 5 | Brown | Male | Hazel |
| 5 | Brown | Male | Green |
| 6 | Red | Female | Brown |
| 6 | Red | Female | Blue |
| 6 | Red | Female | Hazel |
| 6 | Red | Female | Green |

```
7    Red    Male  Brown
7    Red    Male   Blue
7    Red    Male  Hazel
7    Red    Male  Green
```

[72]:
```
# 一列切割为多列
dfc = df[['Hair']].assign(eye_sex = df.Eye+','+df.Sex)
dfc[['Eye', 'Sex']]= dfc['eye_sex'].str.split(',', expand = True)
dfc
```

[72]:

| | Hair | eye_sex | Eye | Sex |
|---|---|---|---|---|
| 0 | Black | Brown,Male | Brown | Male |
| 1 | Brown | Brown,Male | Brown | Male |
| 2 | Red | Brown,Male | Brown | Male |
| 3 | Blond | Brown,Male | Brown | Male |
| 4 | Black | Blue,Male | Blue | Male |
| 5 | Brown | Blue,Male | Blue | Male |
| 6 | Red | Blue,Male | Blue | Male |
| 7 | Blond | Blue,Male | Blue | Male |
| 8 | Black | Hazel,Male | Hazel | Male |
| 9 | Brown | Hazel,Male | Hazel | Male |
| 10 | Red | Hazel,Male | Hazel | Male |
| 11 | Blond | Hazel,Male | Hazel | Male |
| 12 | Black | Green,Male | Green | Male |
| 13 | Brown | Green,Male | Green | Male |
| 14 | Red | Green,Male | Green | Male |
| 15 | Blond | Green,Male | Green | Male |
| 16 | Black | Brown,Female | Brown | Female |
| 17 | Brown | Brown,Female | Brown | Female |
| 18 | Red | Brown,Female | Brown | Female |
| 19 | Blond | Brown,Female | Brown | Female |
| 20 | Black | Blue,Female | Blue | Female |
| 21 | Brown | Blue,Female | Blue | Female |
| 22 | Red | Blue,Female | Blue | Female |
| 23 | Blond | Blue,Female | Blue | Female |
| 24 | Black | Hazel,Female | Hazel | Female |
| 25 | Brown | Hazel,Female | Hazel | Female |

```
26    Red  Hazel,Female  Hazel  Female
27  Blond  Hazel,Female  Hazel  Female
28  Black  Green,Female  Green  Female
29  Brown  Green,Female  Green  Female
30    Red  Green,Female  Green  Female
31  Blond  Green,Female  Green  Female
```

[73]:
```
%%R
# 一行切割为多行
dtr = dt[, paste0(Eye, collapse = ','), keyby = c('Hair', 'Sex')]
dtr[, .(Eye = unlist(strsplit(V1, ','))), by = c('Hair', 'Sex')]
```

```
    Hair    Sex   Eye
 1: Black Female Brown
 2: Black Female  Blue
 3: Black Female Hazel
 4: Black Female Green
 5: Black   Male Brown
 6: Black   Male  Blue
 7: Black   Male Hazel
 8: Black   Male Green
 9: Blond Female Brown
10: Blond Female  Blue
11: Blond Female Hazel
12: Blond Female Green
13: Blond   Male Brown
14: Blond   Male  Blue
15: Blond   Male Hazel
16: Blond   Male Green
17: Brown Female Brown
18: Brown Female  Blue
19: Brown Female Hazel
20: Brown Female Green
21: Brown   Male Brown
22: Brown   Male  Blue
23: Brown   Male Hazel
24: Brown   Male Green
```

```
25:   Red Female Brown
26:   Red Female  Blue
27:   Red Female Hazel
28:   Red Female Green
29:   Red   Male Brown
30:   Red   Male  Blue
31:   Red   Male Hazel
32:   Red   Male Green
     Hair    Sex   Eye
```

[74]:
```
%%R
# 一列切割为多列
dtc = dt[, .(Hair, eye_sex = paste(Eye, Sex, sep = ','))]
dtc[, c('Eye', 'Sex') := tstrsplit(eye_sex, ',')][]
```

```
     Hair      eye_sex   Eye    Sex
 1: Black   Brown,Male Brown   Male
 2: Brown   Brown,Male Brown   Male
 3:   Red   Brown,Male Brown   Male
 4: Blond   Brown,Male Brown   Male
 5: Black    Blue,Male  Blue   Male
 6: Brown    Blue,Male  Blue   Male
 7:   Red    Blue,Male  Blue   Male
 8: Blond    Blue,Male  Blue   Male
 9: Black   Hazel,Male Hazel   Male
10: Brown   Hazel,Male Hazel   Male
11:   Red   Hazel,Male Hazel   Male
12: Blond   Hazel,Male Hazel   Male
13: Black   Green,Male Green   Male
14: Brown   Green,Male Green   Male
15:   Red   Green,Male Green   Male
16: Blond   Green,Male Green   Male
17: Black Brown,Female Brown Female
18: Brown Brown,Female Brown Female
19:   Red Brown,Female Brown Female
20: Blond Brown,Female Brown Female
21: Black  Blue,Female  Blue Female
```

```
22: Brown  Blue,Female  Blue Female
23:   Red  Blue,Female  Blue Female
24: Blond  Blue,Female  Blue Female
25: Black Hazel,Female Hazel Female
26: Brown Hazel,Female Hazel Female
27:   Red Hazel,Female Hazel Female
28: Blond Hazel,Female Hazel Female
29: Black Green,Female Green Female
30: Brown Green,Female Green Female
31:   Red Green,Female Green Female
32: Blond Green,Female Green Female
     Hair      eye_sex   Eye    Sex
```

### 0.11 数据框行合并

[76]:
```
# 数据框行切割
dfdict = dict(tuple(df.groupby(['Sex'])))
# or
dflist = [d for _, d in df.groupby(['Sex'])]
dflist
```

[76]:
```
[    rownames   Hair    Eye     Sex  Freq  freq_total
 16        17  Black  Brown  Female    36         313
 17        18  Brown  Brown  Female    66         313
 18        19    Red  Brown  Female    16         313
 19        20  Blond  Brown  Female     4         313
 20        21  Black   Blue  Female     9         313
 21        22  Brown   Blue  Female    34         313
 22        23    Red   Blue  Female     7         313
 23        24  Blond   Blue  Female    64         313
 24        25  Black  Hazel  Female     5         313
 25        26  Brown  Hazel  Female    29         313
 26        27    Red  Hazel  Female     7         313
 27        28  Blond  Hazel  Female     5         313
 28        29  Black  Green  Female     2         313
 29        30  Brown  Green  Female    14         313
```

```
30        31    Red  Green  Female     7         313
31        32  Blond  Green  Female     8         313,
    rownames   Hair    Eye   Sex  Freq  freq_total
0          1  Black  Brown  Male    32         279
1          2  Brown  Brown  Male    53         279
2          3    Red  Brown  Male    10         279
3          4  Blond  Brown  Male     3         279
4          5  Black   Blue  Male    11         279
5          6  Brown   Blue  Male    50         279
6          7    Red   Blue  Male    10         279
7          8  Blond   Blue  Male    30         279
8          9  Black  Hazel  Male    10         279
9         10  Brown  Hazel  Male    25         279
10        11    Red  Hazel  Male     7         279
11        12  Blond  Hazel  Male     5         279
12        13  Black  Green  Male     3         279
13        14  Brown  Green  Male    15         279
14        15    Red  Green  Male     7         279
15        16  Blond  Green  Male     8         279]
```

[77]:
```
# 数据框行合并
df_con = pd.concat(dfdict, axis=0).reset_index(drop=True)
df_con
```

[77]:

| | rownames | Hair | Eye | Sex | Freq | freq_total |
|---|---|---|---|---|---|---|
| 0 | 17 | Black | Brown | Female | 36 | 313 |
| 1 | 18 | Brown | Brown | Female | 66 | 313 |
| 2 | 19 | Red | Brown | Female | 16 | 313 |
| 3 | 20 | Blond | Brown | Female | 4 | 313 |
| 4 | 21 | Black | Blue | Female | 9 | 313 |
| 5 | 22 | Brown | Blue | Female | 34 | 313 |
| 6 | 23 | Red | Blue | Female | 7 | 313 |
| 7 | 24 | Blond | Blue | Female | 64 | 313 |
| 8 | 25 | Black | Hazel | Female | 5 | 313 |
| 9 | 26 | Brown | Hazel | Female | 29 | 313 |
| 10 | 27 | Red | Hazel | Female | 7 | 313 |
| 11 | 28 | Blond | Hazel | Female | 5 | 313 |

```
12        29  Black  Green  Female     2         313
13        30  Brown  Green  Female    14         313
14        31    Red  Green  Female     7         313
15        32  Blond  Green  Female     8         313
16         1  Black  Brown    Male    32         279
17         2  Brown  Brown    Male    53         279
18         3    Red  Brown    Male    10         279
19         4  Blond  Brown    Male     3         279
20         5  Black   Blue    Male    11         279
21         6  Brown   Blue    Male    50         279
22         7    Red   Blue    Male    10         279
23         8  Blond   Blue    Male    30         279
24         9  Black  Hazel    Male    10         279
25        10  Brown  Hazel    Male    25         279
26        11    Red  Hazel    Male     7         279
27        12  Blond  Hazel    Male     5         279
28        13  Black  Green    Male     3         279
29        14  Brown  Green    Male    15         279
30        15    Red  Green    Male     7         279
31        16  Blond  Green    Male     8         279
```

[78]:
```
%%R
# 数据框行切割
dtlist1 = split(dt, by = 'Sex')
# or
dtlist2 = split(dt, list(dt$Sex))
dtlist2
```

```
$Female
   rownames  Hair   Eye    Sex Freq freq_total
1:       17 Black Brown Female   36        313
2:       18 Brown Brown Female   66        313
3:       19   Red Brown Female   16        313
4:       20 Blond Brown Female    4        313
5:       21 Black  Blue Female    9        313
6:       22 Brown  Blue Female   34        313
7:       23   Red  Blue Female    7        313
```

```
 8:       24 Blond  Blue Female   64        313
 9:       25 Black Hazel Female    5        313
10:       26 Brown Hazel Female   29        313
11:       27   Red Hazel Female    7        313
12:       28 Blond Hazel Female    5        313
13:       29 Black Green Female    2        313
14:       30 Brown Green Female   14        313
15:       31   Red Green Female    7        313
16:       32 Blond Green Female    8        313

$Male
    rownames  Hair   Eye  Sex Freq freq_total
 1:        1 Black Brown Male   32        279
 2:        2 Brown Brown Male   53        279
 3:        3   Red Brown Male   10        279
 4:        4 Blond Brown Male    3        279
 5:        5 Black  Blue Male   11        279
 6:        6 Brown  Blue Male   50        279
 7:        7   Red  Blue Male   10        279
 8:        8 Blond  Blue Male   30        279
 9:        9 Black Hazel Male   10        279
10:       10 Brown Hazel Male   25        279
11:       11   Red Hazel Male    7        279
12:       12 Blond Hazel Male    5        279
13:       13 Black Green Male    3        279
14:       14 Brown Green Male   15        279
15:       15   Red Green Male    7        279
16:       16 Blond Green Male    8        279
```

[80]:
```
%%R
# 数据框行合并
dtbind2 = rbindlist(dtlist1)
dtbind2
```

```
    rownames  Hair   Eye    Sex Freq freq_total
 1:        1 Black Brown   Male   32        279
```

```
 2:        2 Brown Brown   Male   53        279
 3:        3   Red Brown   Male   10        279
 4:        4 Blond Brown   Male    3        279
 5:        5 Black  Blue   Male   11        279
 6:        6 Brown  Blue   Male   50        279
 7:        7   Red  Blue   Male   10        279
 8:        8 Blond  Blue   Male   30        279
 9:        9 Black Hazel   Male   10        279
10:       10 Brown Hazel   Male   25        279
11:       11   Red Hazel   Male    7        279
12:       12 Blond Hazel   Male    5        279
13:       13 Black Green   Male    3        279
14:       14 Brown Green   Male   15        279
15:       15   Red Green   Male    7        279
16:       16 Blond Green   Male    8        279
17:       17 Black Brown Female   36        313
18:       18 Brown Brown Female   66        313
19:       19   Red Brown Female   16        313
20:       20 Blond Brown Female    4        313
21:       21 Black  Blue Female    9        313
22:       22 Brown  Blue Female   34        313
23:       23   Red  Blue Female    7        313
24:       24 Blond  Blue Female   64        313
25:       25 Black Hazel Female    5        313
26:       26 Brown Hazel Female   29        313
27:       27   Red Hazel Female    7        313
28:       28 Blond Hazel Female    5        313
29:       29 Black Green Female    2        313
30:       30 Brown Green Female   14        313
31:       31   Red Green Female    7        313
32:       32 Blond Green Female    8        313
    rownames  Hair   Eye    Sex Freq freq_total
```

## 0.12 数据框列合并

[ ]:
```
df1 = df.sample(n=2).drop('Unnamed: 0', axis=1)
df2 = df.sample(n=3).drop('Unnamed: 0', axis=1)
df3 = df.sample(n=4).drop('Unnamed: 0', axis=1)

# 合并两个数据框
dfmerge2 = pd.merge(
  df1, df2,
  on = ['Hair', 'Eye', 'Sex'],
  how = 'outer'
)
# how: left, right, inner, outer
dfmerge2
```

[ ]:
```
# 合并多个数据框
from functools import reduce
df_merge3 = reduce(
  lambda x,y: pd.merge(
    x,y,
    on = ['Hair', 'Eye', 'Sex'],
    how = 'outer'
  ),
  [df1, df2, df3]
)
```

[ ]:
```
%%R
dt1 = dt[sample(.N,2)][,V1 := NULL]
dt2 = dt[sample(.N,3)][,V1 := NULL]
dt3 = dt[sample(.N,4)][,V1 := NULL]

# 合并两个数据框
dtmerge2 = merge(
  dt1, dt2,
  by = c('Hair', 'Eye', 'Sex'),
  all = TRUE
)
```

```
# all, all.x, all.y: TRUE, FALSE

# 合并多个数据框
dtmerge3 = Reduce(
  function(x,y) merge(
    x,y,
    by = c('Hair', 'Eye', 'Sex'),
    all = TRUE
  ),
  list(dt1, dt2, dt3)
)
```