

Data manipulation I

KY Park
2018 3 29

dplyr

dplyr package

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.3
```

- `select()`
- `filter()`
- `group_by()`
- `summarise()`
- `arrange()`
- `join()`
- `mutate()`

Sample data

```
# data()
```

```
data("infern")
```

```
?infern
```

```
## starting httpd help server ... done
```

```
names(infert)
```

```
## [1] "education" "age" "parity" "induced"  
## [5] "case" "spontaneous" "stratum" "pooled.stratum"
```

Choose column with select()

```
select(infert, age, education) %>% head()
```

```
##   age education
## 1  26    0-5yrs
## 2  42    0-5yrs
## 3  39    0-5yrs
## 4  34    0-5yrs
## 5  35    6-11yrs
## 6  36    6-11yrs
```

```
select(infert, -age) %>% head()
```

```
##   education parity induced case spontaneous stratum pooled.stratum
## 1    0-5yrs     6       1    1           2         1             3
## 2    0-5yrs     1       1    1           0         2             1
## 3    0-5yrs     6       2    1           0         3             4
## 4    0-5yrs     4       2    1           0         4             2
## 5    6-11yrs     3       1    1           1         5            32
## 6    6-11yrs     4       2    1           1         6            36
```

Extract unique values with distinct()

```
select(infert, education) %>% distinct()
```

```
##   education
```

```
## 1    0-5yrs
```

```
## 2    6-11yrs
```

```
## 3   12+ yrs
```

Choose row with filter()

```
filter(infert, age > 35) %>% head()
```

```
##   education age parity induced case spontaneous stratum pooled.stratum
## 1    0-5yrs  42     1      1      1              0        2            1
## 2    0-5yrs  39     6      2      1              0        3            4
## 3    6-11yrs 36     4      2      1              1        6           36
## 4    6-11yrs 37     4      2      1              1       12           37
## 5    6-11yrs 44     1      0      1              1       20           17
## 6    6-11yrs 40     1      0      1              1       21           14
```


Choose row with filter() 2

```
filter(infert, education == "0-5yrs") %>% nrow()
```

```
## [1] 12
```

Choose row with filter() 3

```
filter(infert, age > 35, education == "0-5yrs") %>% head()
```

```
##   education age parity induced case spontaneous stratum pooled.stratum
## 1    0-5yrs  42     1       1     1           0         2           1
## 2    0-5yrs  39     6       2     1           0         3           4
## 3    0-5yrs  42     1       0     0           0         2           1
## 4    0-5yrs  39     6       2     0           0         3           4
## 5    0-5yrs  42     1       0     0           0         2           1
## 6    0-5yrs  39     6       2     0           0         3           4
```

Change order with arrange()

```
arrange(infert, age, desc(parity))
```

##	education	age	parity	induced	case	spontaneous	stratum	pooled.stratum
## 1	6-11yrs	21	1	0	1	1	9	5
## 2	12+ yrs	21	1	0	1	1	67	39
## 3	6-11yrs	21	1	0	0	1	9	5
## 4	12+ yrs	21	1	0	0	1	67	39
## 5	6-11yrs	21	1	1	0	0	9	5
## 6	12+ yrs	21	1	0	0	0	67	39
## 7	6-11yrs	23	1	0	1	0	7	6
## 8	12+ yrs	23	1	0	1	1	83	40
## 9	6-11yrs	23	1	0	0	0	7	6
## 10	12+ yrs	23	1	0	0	1	83	40
## 11	6-11yrs	23	1	0	0	0	7	6
## 12	12+ yrs	23	1	0	0	1	83	40
## 13	12+ yrs	24	3	1	1	2	51	56
## 14	12+ yrs	24	3	2	0	1	51	56
## 15	12+ yrs	24	3	2	0	0	51	56
## 16	6-11yrs	25	3	2	1	1	19	28
## 17	6-11yrs	25	3	0	0	1	19	28

```
db_test <- data.frame(a=1:5, b=c("a", "b", "a", "a", "b"))  
db_test
```

```
##   a b  
## 1 1 a  
## 2 2 b  
## 3 3 a  
## 4 4 a  
## 5 5 b
```

```
group_by(db_test, b)
```

```
## # A tibble: 5 x 2
```

```
## # Groups:   b [2]
```

```
##       a       b
```

```
##   <int> <fctr>
```

```
## 1     1     a
```

```
## 2     2     b
```

```
## 3     3     a
```

```
## 4     4     a
```

```
## 5     5     b
```

```
summarize(db_test, meanOfA = mean(a))
```

```
##   meanOfA
```

```
## 1       3
```

```
summarize(db_test, minOfA = min(a), meanOfA = mean(a), sdOfA = sd(a))
```

```
##   minOfA meanOfA   sdOfA  
## 1      1      3 1.581139
```

```
group_by(db_test, b) %>% summarize(minOfA = min(a), meanOfA = mean(a), sdOfA = sd(a))
```

```
## # A tibble: 2 x 4  
##       b minOfA meanOfA   sdOfA  
##   <fctr> <dbl>   <dbl>   <dbl>  
## 1     a     1 2.666667 1.527525  
## 2     b     2 3.500000 2.121320
```

```
group_by(infert, education) %>%  
  summarize(mean(age), min(age), max(age))
```

```
## # A tibble: 3 x 4
```

```
##   education `mean(age)` `min(age)` `max(age)`  
##   <fctr>      <dbl>      <dbl>      <dbl>  
## 1  0-5yrs    35.25000      26        42  
## 2  6-11yrs   32.85000      21        44  
## 3 12+ yrs    29.72414      21        38
```

combining data using dplyr

a

x1	x2
A	1
B	2
C	3

+

b

x1	x3
A	T
B	F
D	T

=

Mutating Joins

x1	x2	x3
A	1	T
B	2	F
C	3	NA

dplyr::left_join(a, b, by = "x1")
Join matching rows from b to a.

x1	x3	x2
A	T	1
B	F	2
D	T	NA

dplyr::right_join(a, b, by = "x1")
Join matching rows from a to b.

x1	x2	x3
A	1	T
B	2	F

dplyr::inner_join(a, b, by = "x1")
Join data. Retain only rows in both sets.

x1	x2	x3
A	1	T
B	2	F
C	3	NA
D	NA	T

dplyr::full_join(a, b, by = "x1")
Join data. Retain all values, all rows.

Filtering Joins

x1	x2
A	1
B	2

dplyr::semi_join(a, b, by = "x1")
All rows in a that have a match in b.

x1	x2
C	3

dplyr::anti_join(a, b, by = "x1")
All rows in a that do not have a match in b.

y		z	
x1	x2	x1	x2
A	1	B	2
B	2	C	3
C	3	D	4

+

=

Set Operations

x1	x2
B	2
C	3

dplyr::intersect(y, z)

Rows that appear in both y and z.

x1	x2
A	1
B	2
C	3
D	4

dplyr::union(y, z)

Rows that appear in either or both y and z.

x1	x2
A	1

dplyr::setdiff(y, z)

Rows that appear in y but not z.

Binding

x1	x2
A	1
B	2
C	3
B	2
C	3
D	4

dplyr::bind_rows(y, z)

Append z to y as new rows.

x1	x2	x1	x2
A	1	B	2
B	2	C	3
C	3	D	4

dplyr::bind_cols(y, z)

Append z to y as new columns.

Caution: matches rows by position.

Mutate()

```
meanage <- mean(infert$age)
mutate(infert, age_centering=age-mean(age)) %>%
  select(age, age_centering) %>% head()
```

```
##   age age_centering
## 1  26    -5.504032
## 2  42    10.495968
## 3  39     7.495968
## 4  34     2.495968
## 5  35     3.495968
## 6  36     4.495968
```

Exercise 1

Explain the following code

```
infert %>%  
  filter(education == "0-5yrs") %>%  
  arrange(age) %>%  
  mutate(age_centering=age-mean(age)) %>%  
  select(education, age, age_centering)
```

```
##      education age age_centering  
## 1      0-5yrs  26         -9.25  
## 2      0-5yrs  26         -9.25  
## 3      0-5yrs  26         -9.25  
## 4      0-5yrs  34         -1.25  
## 5      0-5yrs  34         -1.25  
## 6      0-5yrs  34         -1.25  
## 7      0-5yrs  39          3.75  
## 8      0-5yrs  39          3.75  
## 9      0-5yrs  39          3.75  
## 10     0-5yrs  42          6.75  
## 11     0-5yrs  42          6.75  
## 12     0-5yrs  42          6.75
```

- Make two data frame named db_a (variable: ID, age) and db_b (variable: ID, parity)
- try the following functions: left_join(), right_join(), inner_join(), and full_join()

tidyr

The effect of drug on heart rate

```
messy <- data.frame(  
  ID = c("StudyID_1", "StudyID_2", "StudyID_3", "StudyID_4"),  
  DrugA = c(56, 67, 60, 40),  
  DrugB = c(70, 102, 90, 83)  
)  
messy
```

##	ID	DrugA	DrugB
## 1	StudyID_1	56	70
## 2	StudyID_2	67	102
## 3	StudyID_3	60	90
## 4	StudyID_4	40	83

Each column is a variable.

Each row is an observation.

gather

```
gather(messy, Drug, HeartRate, DrugA:DrugB)
```

```
##           ID  Drug HeartRate
## 1 StudyID_1 DrugA         56
## 2 StudyID_2 DrugA         67
## 3 StudyID_3 DrugA         60
## 4 StudyID_4 DrugA         40
## 5 StudyID_1 DrugB         70
## 6 StudyID_2 DrugB        102
## 7 StudyID_3 DrugB         90
## 8 StudyID_4 DrugB         83
```