Becoming a data ninja with dplyr

Devin Pastoor
Center for Translational Medicine
University of Maryland, School of Pharmacy

major data manipulation verbs

Verb	Usage
filter	Keep matching row criteria
summarize	reduces summary values calculated
mutate	add new variables to existing data frame
select	select columns by name
arrange	reorder rows

```
df <- data.frame(ID = 1:5,
GENDER = c("MALE", "MALE", "FEMALE", "MALE", "FEMALE"),
WT = c(70, 76, 60, 64, 68))
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

Verb	Usage
filter	Keep matching row criteria
summarize	reduces summary values calculated
mutate	add new variables to existing data frame
select	select columns by name
arrange	reorder rows

```
filter(df, GENDER == "FEMALE")
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT
3	FEMALE	60
5	FEMALE	68

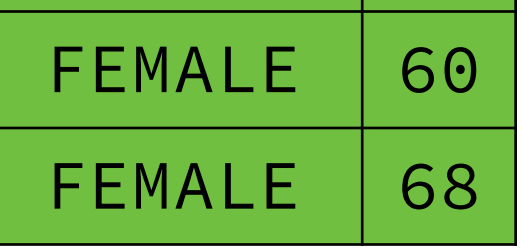
common dplyr filter (subset) operators

operator	meaning
==, !=	equal, not equal
>, >=	greater than, greater than or equal to
<, <=	less than, less than or equal to
is.na, !is.na	is NA, not NA
!duplicated	only first value
%in%	in specified values

filter seperator	base equivalent	meaning
,	&	and
|	|	or

```
filter(df, ID %in% c(1, 3, 5))
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68



ID	GENDER	WT
1	MALE	70
3	FEMALE	60
5	FEMALE	68

```
filter(df, GENDER == "MALE", WT > 70)
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT
2	MALE	76

```
filter(df, GENDER == "FEMALE" | WT < 70)
```

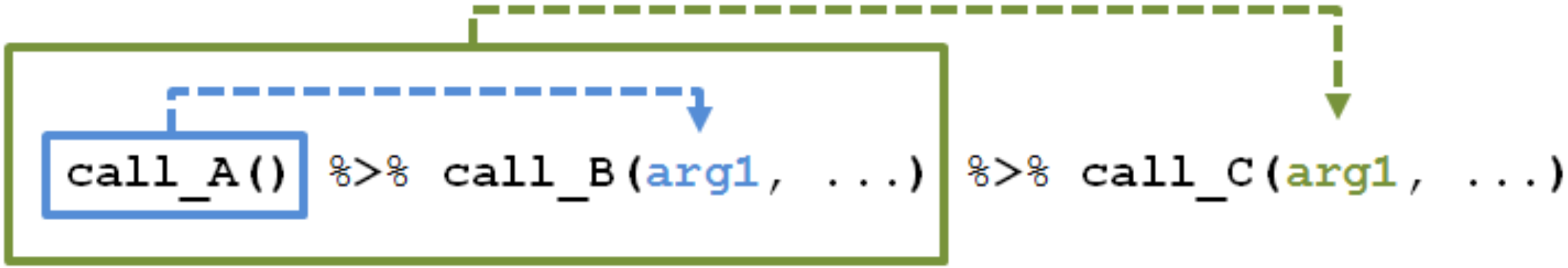
ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT
3	FEMALE	60
4	MALE	64
5	FEMALE	68

Your Turn

chaining and grouped operations

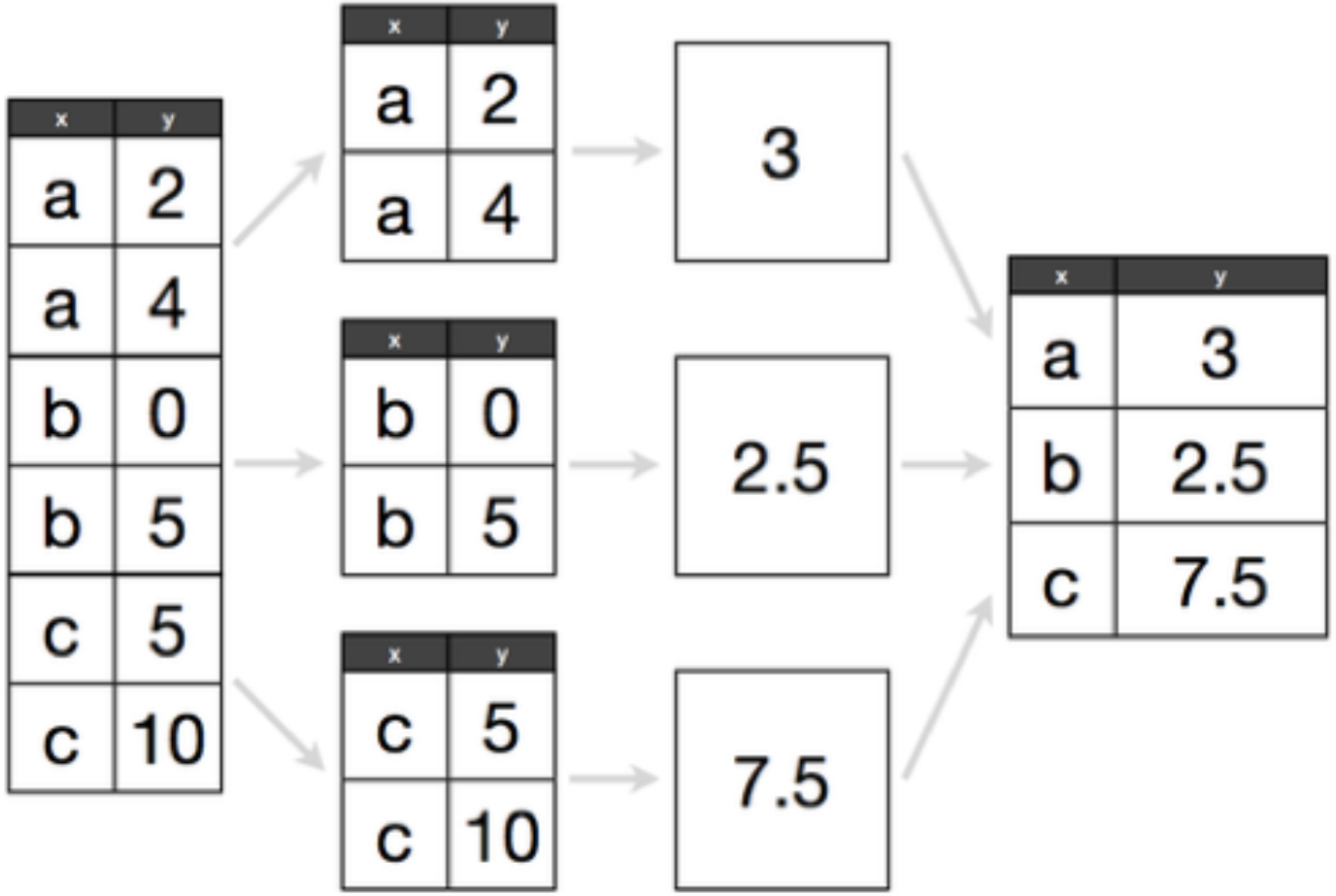
chaining with %>%



%>% is pronounced ‘then’

```
> Theoph %>% filter(!duplicated(Subject)) %>% head()
  Subject   Wt Dose Time conc
1       1 79.6 4.02    0 0.74
2       2 72.4 4.40    0 0.00
3       3 70.5 4.53    0 0.00
4       4 72.7 4.40    0 0.00
5       5 54.6 5.86    0 0.00
6       6 80.0 4.00    0 0.00
```

split-apply-combine -> group_by



http://www.gregreda.com/2013/10/26/working-with-pandas-g

Verb	Usage
filter	Keep matching row criteria
summarize	reduces summary values calculated
mutate	add new variables to existing data frame
select	select columns by name
arrange	reorder rows

```
df %>% summarize(meanWT = mean(WT))
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

meanWT
67.6

```
summarize(df, meanWT = mean(WT))
```

```
df %>% group_by(GENDER) %>%
    summarize(meanWT = mean(WT))
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

GENDER	meanWT
MALE	70
FEMALE	64

```
df %>% group_by(GENDER) %>%
    summarize(meanWT = mean(WT), n = n())
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

GENDER	meanWT	n
MALE	70	3
FEMALE	64	2

Verb	Usage
filter	Keep matching row criteria
summarize	reduces summary values calculated
mutate	add new variables to existing data frame
select	select columns by name
arrange	reorder rows

```
df %>% mutate(meanWT = mean(WT))
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT	meanWT
1	MALE	70	67.6
2	MALE	76	67.6
3	FEMALE	60	67.6
4	MALE	64	67.6
5	FEMALE	68	67.6

```
df %>% group_by(GENDER) %>%
 mutate(meanWT = mean(WT))
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT	meanWT
1	MALE	70	70
2	MALE	76	70
3	FEMALE	60	64
4	MALE	64	70
5	FEMALE	68	64

```
df %>% group_by(GENDER) %>%
 mutate(meanWT = mean(WT),
                mWT_LB = meanWT*2.2)
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT	meanWT	mWT_LB
1	MALE	70	70	154
2	MALE	76	70	154
3	FEMALE	60	64	140.8
4	MALE	64	70	154
5	FEMALE	68	64	140.8

```
df %>%
 mutate(ISM = ifelse(GENDER == "MALE", 1, 0))
```

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT	ISM
1	MALE	70	1
2	MALE	76	1
3	FEMALE	60	0
4	MALE	64	1
5	FEMALE	68	0

Verb	Usage
filter	Keep matching row criteria
summarize	reduces summary values calculated
mutate	add new variables to existing data frame
select	select columns by name
arrange	reorder rows

df2 %>% select(ID, WT)

ID	GENDER	WT	meanWT
1	MALE	70	67.6
2	MALE	76	67.6
3	FEMALE	60	67.6
4	MALE	64	67.6
5	FEMALE	68	67.6

ID	WT
1	70
2	76
3	60
4	64
5	68

df2 %>% select(GENDER:meanWT)

ID	GENDER	WT	meanWT
1	MALE	70	67.6
2	MALE	76	67.6
3	FEMALE	60	67.6
4	MALE	64	67.6
5	FEMALE	68	67.6

GENDER	WT	meanWT
MALE	70	67.6
MALE	76	67.6
FEMALE	60	67.6
MALE	64	67.6
FEMALE	68	67.6

df %>% select(<function>(<values>))

df with the following columns:

WEIGHT	WEIGHT_KG	MEAN_WEIGHT	OCC1	OCC2	OCC3	OCC4	HEIGHT

function	meaning	example	columns selected
starts_with	names start with	starts_with("WEIGHT"))	WEIGHT, WEIGHT_KG
ends_with	names ends with	ends_with("GHT")	WEIGHT, MEAN_WEIGHT, HEIGHT
contains	names contains	contains("EI")	WEIGHT, WEIGHT_KG, MEAN_WEIGHT, HEIGHT
matches	regular expression matching	matches("_")	WEIGHT_KG, MEAN_WEIGHT
num_range	specify range of columns with consistent names with numeric suffix	num_range("OCC",1:3)	OCC1, OCC2, OCC3

```
test_select <-
data.frame("WEIGHT" =0, "WEIGHT_KG"=0,
       "MEAN_WEIGHT"=0, "OCC1"=0, "OCC2"=0,
       "OCC3"=0, "OCC4"=0, "HEIGHT"=0)
```

```
test_select %>% select(starts_with("WEIGHT"))
test_select %>% select(matches("_"))
test_select %>% select(num_range("OCC", 1:3)
```

```
df2 %>% select(ID, WEIGHT = WT)
```

ID	GENDER	WT	meanWT
1	MALE	70	67.6
2	MALE	76	67.6
3	FEMALE	60	67.6
4	MALE	64	67.6
5	FEMALE	68	67.6

ID	WEIGHT
1	70
2	76
3	60
4	64
5	68

Verb	Usage
filter	Keep matching row criteria
summarize	reduces summary values calculated
mutate	add new variables to existing data frame
select	select columns by name
arrange	reorder rows

df %>% arrange(WT)

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT
3	FEMALE	60
4	MALE	64
5	FEMALE	68
1	MALE	70
2	MALE	76

lowest weight

highest weight

df %>% arrange(desc(WT))

ID	GENDER	WT
1	MALE	70
2	MALE	76
3	FEMALE	60
4	MALE	64
5	FEMALE	68

ID	GENDER	WT
2	MALE	76
1	MALE	70
5	FEMALE	68
4	MALE	64
3	FEMALE	60

highest weight

lowest weight

```
dosing_df <- data.frame(ID = 1:2, TIME =0, AMT = 100, MDV = 1)
sample_df <- data.frame(expand.grid(ID = 1:2, TIME = seq(0, 2, 1),
                                    AMT = 0, MDV = 0)
df3 <- rbind(sample_df, dosing_df)
```

* expand.grid is a very handy function for generating permutations

* MDV = missing dependent variable - a nonmem-style flag column

ID	TIME	AMT	MDV
1	0	0	0
2	0	0	0
1	1	0	0
2	1	0	0
1	2	0	0
2	2	0	0
1	0	100	1
2	0	100	1

df3

df3 %>% arrange(ID, TIME)

ID	TIME	AMT	MDV
1	0	0	0
2	0	0	0
1	1	0	0
2	1	0	0
1	2	0	0
2	2	0	0
1	0	100	1
2	0	100	1

ID	TIME	AMT	MDV
1	0	0	0
1	0	100	1
1	1	0	0
1	2	0	0
2	0	0	0
2	0	100	1
2	1	0	0
2	2	0	0

```
df3 %>% arrange(ID, TIME, desc(MDV))
```

ID	TIME	AMT	MDV
1	0	0	0
2	0	0	0
1	1	0	0
2	1	0	0
1	2	0	0
2	2	0	0
1	0	100	1
2	0	100	1

ID	TIME	AMT	MDV
1	0	100	1
1	0	0	0
1	1	0	0
1	2	0	0
2	0	100	1
2	0	0	0
2	1	0	0
2	2	0	0

minor data manipulation verbs

Verb	Usage
transmute	similar to mutate, however all columns not involved in calculation (either grouping or explicit calculations) are dropped
slice	return certain rows by number, must be INTEGER

(some) additional functions

Let’s take a look

- top_n
- all.equal
- lappy_cluster
- **distinct**
- **group_indices**
- **group_size**
- n_groups
- lead/lag
- sample
- tally*
- first
- last

dplyr joins

Join	Usage
inner_join	return all rows from x where there are matching values in y, and all columns from x and y.
left_join	return all rows from x, and all columns from x and y.
semi_join	return all rows from x where there are matching values in y, keeping just columns from x.
anti_join	return all rows from x where there are not matching values in y, keeping just columns from x.
full_join	returns all rows and columns from x and y, with NA values for non-matching values from either.

* A semi join differs from an inner join because an inner join will return one row of x for each matching row of y, where a semi join will never duplicate rows of x

```
idtime <- data.frame(expand.grid(ID = as.numeric(1:3),
TIME = c(0,1))) %>% arrange(ID)
idwt <- data.frame(ID = c(1, 2, 4), WT = c(70, 80, 75))
```

ID	TIME
1	0
1	1
2	0
2	1
3	0
3	1

idtime

ID	WT
1	70
2	80
4	75

idwt

<join>(x_df, y_df)

INNER JOIN

idtime/idwt => in both

ID	TIME
1	0
1	1
2	0
2	1
3	0
3	1

idtime

ID	WT
1	70
2	80
4	75

idwt

inner_join(idtime, idwt)

ID	TIME	WT
1	0	70
1	1	70
2	0	80
2	1	80

inner_join(idwt, idtime)

ID	WT	TIME
1	70	0
1	70	1
2	80	0
2	80	1

LEFT JOIN

idtime/idwt => in both

ID	TIME
1	0
1	1
2	0
2	1
3	0
3	1

idtime

ID	WT
1	70
2	80
4	75

idwt

left_join(idtime, idwt)

ID	TIME	WT
1	0	70
1	1	70
2	0	80
2	1	80
3	0	NA
3	1	NA

left_join(idwt, idtime)

ID	WT	TIME
1	70	0
1	70	1
2	80	0
2	80	1
4	75	NA

SEMI JOIN

idtime/idwt => in both

ID	TIME
1	0
1	1
2	0
2	1
3	0
3	1

idtime

ID	WT
1	70
2	80
4	75

idwt

semi_join(idtime, idwt)

ID	TIME
1	0
1	1
2	0
2	1

ID	WT
1	70
2	80

semi_join(idwt, idtime)

ANTI JOIN

idtime/idwt => in both

ID	TIME
1	0
1	1
2	0
2	1
3	0
3	1

idtime

ID	WT
1	70
2	80
4	75

idwt

anti_join(idtime, idwt)

ID	TIME
3	0
3	1

anti_join(idwt, idtime)

ID	WT
4	75

FULL JOIN

idtime/idwt => in both

ID	TIME
1	0
1	1
2	0
2	1
3	0
3	1

idtime

ID	WT
1	70
2	80
4	75

idwt

full_join(idtime, idwt)

ID	TIME	WT
1	0	70
1	1	70
2	0	80
2	1	80
3	0	NA
3	1	NA
4	NA	75

full_join(idwt, idtime)

ID	WT	TIME
1	70	0
1	70	1
2	80	0
2	80	1
3	NA	0
3	NA	1
4	75	NA