Workshop Series:
Reusable Research Data Made Shiny

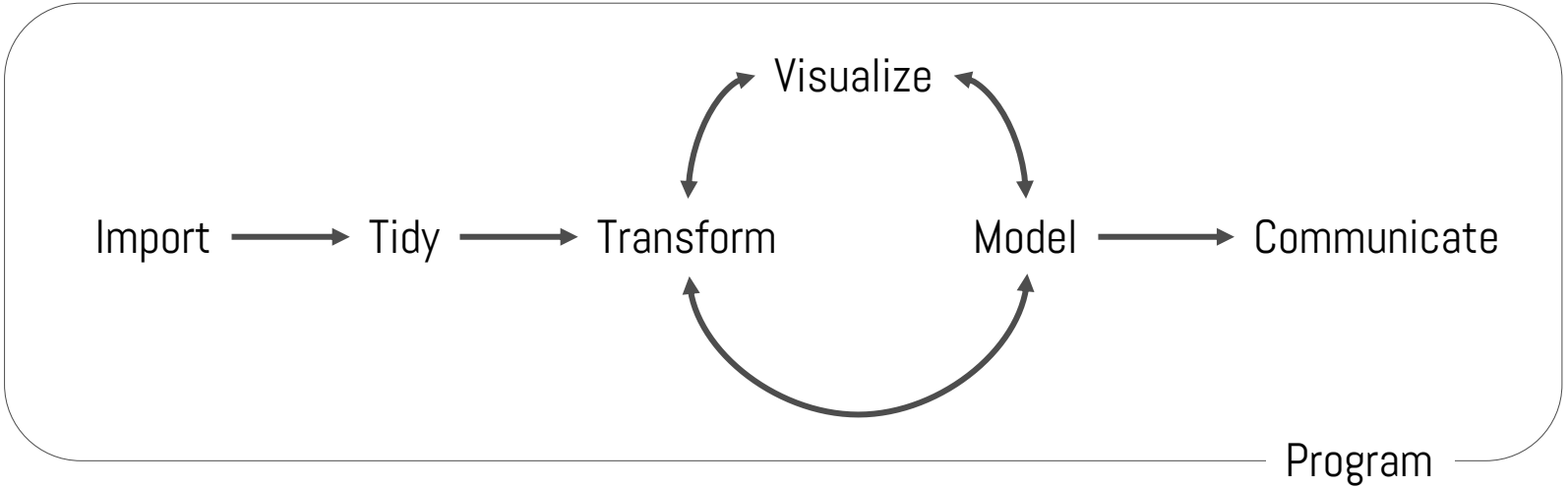
Ontario Dairy Research Centre | Online
February 21st - 24th, 2023

University of Guelph
Improve Life.

Compute Ontario

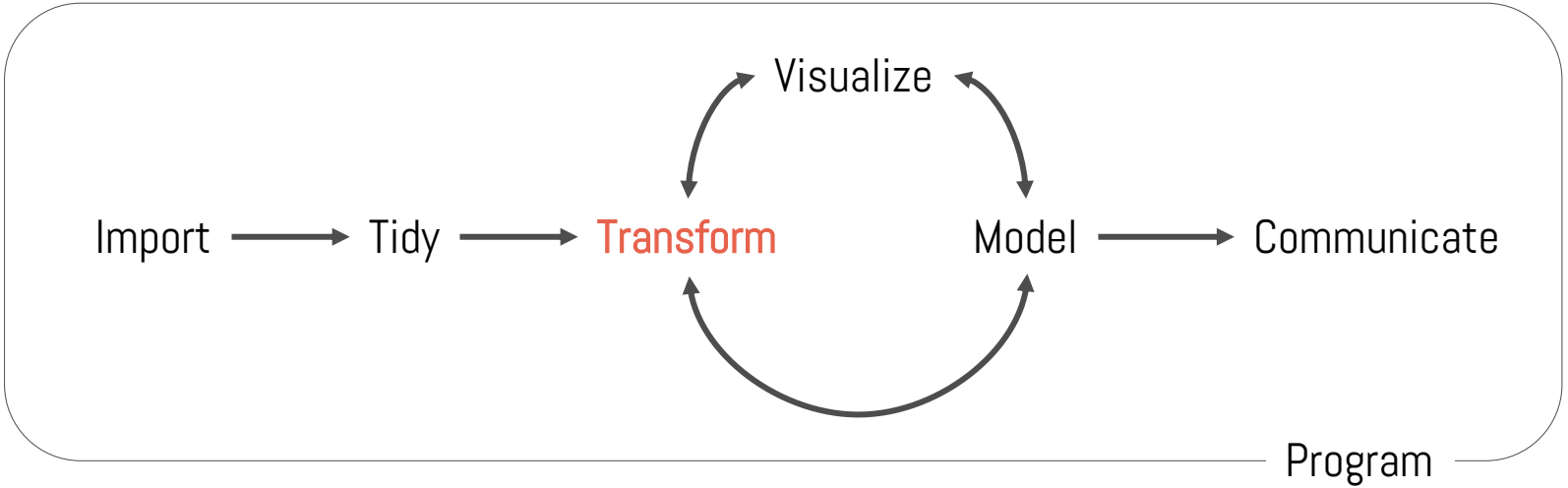
ADC

Basic data workflow



Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

Basic data workflow



Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

Toy data

Let's look at our toy data environmental_data.csv

head(env_data)

```
# A tibble: 124,744 × 6
# Groups:   date, time, barn [62,372]
   date       time   barn      location    rh  temp
   <date>     <time> <chr>     <chr>    <dbl> <dbl>
 1 2022-01-01 11'58" lactating inside      74  7.74
 2 2022-01-01 11'58" lactating outside     87  2.5
 3 2022-01-01 11'58" sp_needs  inside      78 10.1
 4 2022-01-01 11'58" sp_needs  outside     87  2.5
 5 2022-01-01 26'58" lactating inside      74  8.31
 6 2022-01-01 26'58" lactating outside     87  2.5
 7 2022-01-01 26'58" sp_needs  inside      77  9.95
 8 2022-01-01 26'58" sp_needs  outside     87  2.5
 9 2022-01-01 41'58" lactating inside      74  8.89
10 2022-01-01 41'58" lactating outside     87  2.5
# … with 124,734 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

Toy data

Let's take a closer look at our toy data:
skim(env_data)

```
── Data Summary ────────────────────────
                           Values
Name                       env_data
Number of rows             124744
Number of columns          6
_______________________
Column type frequency:
  character                2
  Date                     1
  difftime                 1
  numeric                  2
_______________________
Group variables            None
```

Toy data

Let's take a closer look at our toy data:
skim(env_data)

```
── Variable type: character ──────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 barn                  0             1   8   9     0        2          0
2 location              0             1   6   7     0        2          0

── Variable type: Date ───────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min        max        median     n_unique
1 date                  0             1 2022-01-01 2022-12-31 2022-07-22      325

── Variable type: difftime ───────────────────────────────────────────────────
  skim_variable n_missing complete_rate min      max        median       n_unique
1 time                  0             1 718 secs 86219 secs 43468.5 secs      362

── Variable type: numeric ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean    sd  p0   p25  p50 p75 p100 hist
1 rh                    8          1.00 69.7 14.0   11 64    74    79 93   ▁▂▇
2 temp                  8          1.00 12.2  9.70 -24  7.39 12.4  19 39.2 ▁▃▇
```

Ultimate question

Are the temperatures inside the barns milder than outside?

i.e., warmer in the winter and colder in the summer

Isolating data

Extract variables with **select()**

Extract observations with **filter()**

Arrange/Sort observations with **arrange()**


filter()

filter()

Extract rows that meet logical criteria

filter(.data, ...)

.data	dataframe to transform
...	one or more logical tests (filter returns each row for which the test is TRUE)

filter()

Extract rows that meet logical criteria

filter(env_data, barn == "lactating")

```
# A tibble: 124,744 × 6
# Groups:   date, time, barn [62,372]
   date       time     barn      location    rh  temp
   <date>     <time>   <chr>     <chr>    <dbl> <dbl>
 1 2022-01-01 11'58"   lactating inside      74  7.74
 2 2022-01-01 11'58"   lactating outside     87  2.5
 3 2022-01-01 11'58"   sp_needs  inside      78 10.1
 4 2022-01-01 11'58"   sp_needs  outside     87  2.5
 5 2022-01-01 26'58"   lactating inside      74  8.31
 6 2022-01-01 26'58"   lactating outside     87  2.5
 7 2022-01-01 26'58"   sp_needs  inside      77  9.95
 8 2022-01-01 26'58"   sp_needs  outside     87  2.5
 9 2022-01-01 41'58"   lactating inside      74  8.89
10 2022-01-01 41'58"   lactating outside     87  2.5
```

```
# A tibble: 62,372 × 6
   date       time     barn      location    rh  temp
   <date>     <time>   <chr>     <chr>    <dbl> <dbl>
 1 2022-01-01 00:11:58 lactating inside      74  7.74
 2 2022-01-01 00:11:58 lactating outside     87  2.5
 3 2022-01-01 00:26:58 lactating inside      74  8.31
 4 2022-01-01 00:26:58 lactating outside     87  2.5
 5 2022-01-01 00:41:58 lactating inside      74  8.89
 6 2022-01-01 00:41:58 lactating outside     87  2.5
 7 2022-01-01 00:56:58 lactating inside      74  8.28
 8 2022-01-01 00:56:58 lactating outside     87  2.6
 9 2022-01-01 01:11:58 lactating inside      75  8.6
10 2022-01-01 01:11:58 lactating outside     87  2.8
```

Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

filter()

Extract rows that meet logical criteria

filter(env_data, barn == "lactating")

== tests if equal
(returns TRUE or FALSE)

= used to set things
(returns nothing)

```
# A tibble: 124,744 × 6
# Groups:   date, time, barn [62,372]
   date       time     barn      location    rh  temp
   <date>     <time>   <chr>     <chr>    <dbl> <dbl>
 1 2022-01-01 11'58"   lactating inside      74  7.74
 2 2022-01-01 11'58"   lactating outside     87  2.5
 3 2022-01-01 11'58"   sp_needs  inside      78 10.1
 4 2022-01-01 11'58"   sp_needs  outside     87  2.5
 5 2022-01-01 26'58"   lactating inside      74  8.31
 6 2022-01-01 26'58"   lactating outside     87  2.5
 7 2022-01-01 26'58"   sp_needs  inside      77  9.95
 8 2022-01-01 26'58"   sp_needs  outside     87  2.5
 9 2022-01-01 41'58"   lactating inside      74  8.89
10 2022-01-01 41'58"   lactating outside     87  2.5
```

```
# A tibble: 62,372 × 6
   date       time     barn      location    rh  temp
   <date>     <time>   <chr>     <chr>    <dbl> <dbl>
 1 2022-01-01 00:11:58 lactating inside      74  7.74
 2 2022-01-01 00:11:58 lactating outside     87  2.5
 3 2022-01-01 00:26:58 lactating inside      74  8.31
 4 2022-01-01 00:26:58 lactating outside     87  2.5
 5 2022-01-01 00:41:58 lactating inside      74  8.89
 6 2022-01-01 00:41:58 lactating outside     87  2.5
 7 2022-01-01 00:56:58 lactating inside      74  8.28
 8 2022-01-01 00:56:58 lactating outside     87  2.6
 9 2022-01-01 01:11:58 lactating inside      75  8.6
10 2022-01-01 01:11:58 lactating outside     87  2.8
```

Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

Logical tests

x < y	Less than
x > y	Greater than
x == y	Equals to
x <= y	Less than or equal to
x >= y	Grater than or equal to
x != y	Not equal to

x %in% y	x is at least one of y
is.na(x)	Is NA
!is.na(x)	Is not NA
a & b	and
a | b	or

Your turn!

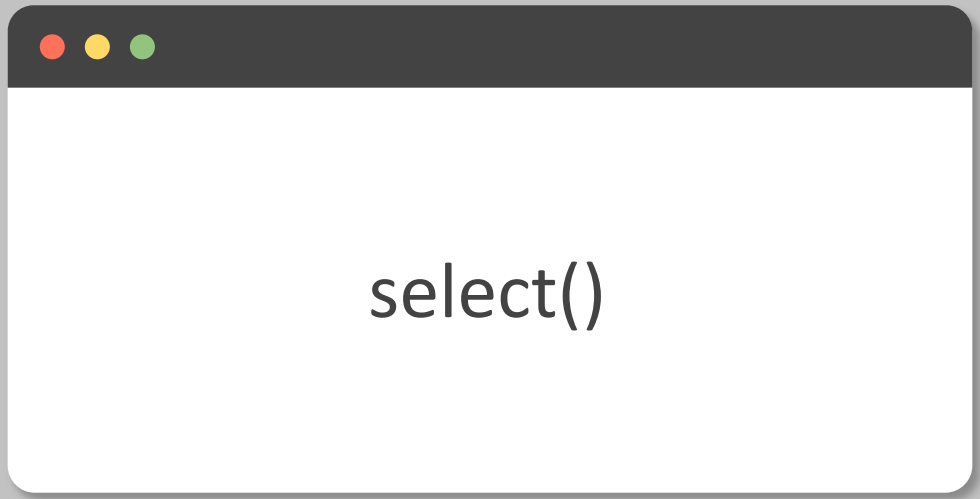
Use the logical operators to manipulate our env_data to show observations that are:

1. From inside the barns
2. Above 30C
3. Between 0 and 10C inside the Special Needs barn
4. From June 3, 2022, or June 4, 2022

Your turn!

Use the logical operators to manipulate our env_data to show observations that are:

1. From inside the barn
    a. filter(env_data, location == "inside")
2. Above 30C
    a. filter(env_data, temp > 30)
3. Bonus: Between 0 and 10C inside the Special Needs barn
    a. filter(env_data, temp >= 0, temp <= 10, location == "inside", barn == "sp_needs")
4. From June 4, 2022, or June 5, 2022
    a. filter(env_data, date == "2022-06-04" | date == "2022-06-05")


select()

select()

Extract columns by name

select(.data, ...)

.data	dataframe to transform
...	name(s) of columns to extract, or a select helper function

select()

Only show records for temperature

select(env_data, date, time, barn, location, temp)

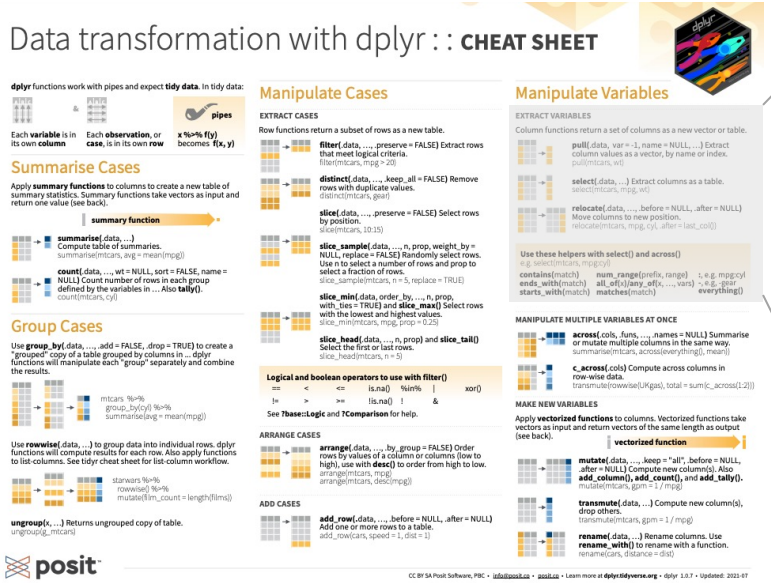
```
# A tibble: 124,744 × 5
   date       time   barn      location  temp
   <date>     <time> <chr>     <chr>    <dbl>
 1 2022-01-01 11'58" lactating inside    7.74
 2 2022-01-01 11'58" lactating outside   2.5
 3 2022-01-01 11'58" sp_needs  inside   10.1
 4 2022-01-01 11'58" sp_needs  outside   2.5
 5 2022-01-01 26'58" lactating inside    8.31
 6 2022-01-01 26'58" lactating outside   2.5
 7 2022-01-01 26'58" sp_needs  inside    9.95
 8 2022-01-01 26'58" sp_needs  outside   2.5
 9 2022-01-01 41'58" lactating inside    8.89
10 2022-01-01 41'58" lactating outside   2.5
```

select() helpers

1. Select a range of columns (:)
   a. select(env_data, date:location)
2. Select every column but (-)
   a. select(env_data, -rh)
3. Select columns that start with … (starts_with())
   a. select(env_data, starts_with("t"))
4. Select columns that end with … (ends_with())
   a. select(env_data, ends_with("e")

select() helpers

And a few more!



EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.

pull(.data, var = -1, name = NULL, …**)** Extract column values as a vector, by name or index.
pull(mtcars, wt)

select(.data, …**)** Extract columns as a table.
select(mtcars, mpg, wt)

relocate(.data, …, .before = NULL, .after = NULL**)** Move columns to new position.
relocate(mtcars, mpg, cyl, .after = last_col())

Use these helpers with select() and across()
e.g. select(mtcars, mpg:cyl)

contains(match**)**
ends_with(match**)**
starts_with(match**)**
num_range(prefix, range**)**
all_of(x**)/any_of(**x, …, vars**)**
matches(match**)**
:, e.g. mpg:cyl
-, e.g, -gear
everything()

https://raw.githubusercontent.com/rstudio/cheatsheets/main/data-transformation.pdf

Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

Your turn!

Use filter() and select() to show only relative humidity from inside the Special Needs barn
*hint: feel free to create an intermediate object

Your turn!

Use filter() and select() to show only relative humidity from inside the Special Needs barn
*hint: feel free to create an intermediate object

```
sp_needs <- filter(env_data, barn == "sp_needs", location == "inside")
sp_needs <- select(sp_needs , date, time, rh)
sp_needs
```

```
# A tibble: 31,186 × 3
   date       time        rh
   <date>     <time>   <dbl>
 1 2022-01-01 00:11:58    78
 2 2022-01-01 00:26:58    77
 3 2022-01-01 00:41:58    77
 4 2022-01-01 00:56:58    78
 5 2022-01-01 01:11:58    78
```


arrange()

arrange()

Order rows from smallest to largest values

arrange(.data, ...)

.data	dataframe to transform
...	One or more columns to order by (additional columns will be used as tie breakers)

arrange()

1. Order by **temp**

arrange(env_data, temp)

```
# A tibble: 124,744 × 6
   date       time     barn      location    rh  temp
   <date>     <time>   <chr>     <chr>    <dbl> <dbl>
 1 2022-01-21 08:11:58 sp_needs  outside     77 -24
 2 2022-01-21 08:11:58 lactating outside     77 -23.9
 3 2022-01-29 05:41:58 sp_needs  outside     71 -23.9
 4 2022-01-29 05:41:58 lactating outside     71 -23.8
 5 2022-01-29 07:26:58 sp_needs  outside     72 -23.8
 6 2022-01-29 07:56:58 sp_needs  outside     73 -23.8
 7 2022-01-29 07:56:58 lactating outside     73 -23.7
 8 2022-01-21 05:41:58 lactating outside     76 -23.6
 9 2022-01-21 05:41:58 sp_needs  outside     77 -23.6
10 2022-01-29 07:26:58 lactating outside     72 -23.6
```

2. Order by **temp** descending

arrange(env_data, desc(temp))

```
# A tibble: 124,744 × 6
   date       time     barn      location    rh  temp
   <date>     <time>   <chr>     <chr>    <dbl> <dbl>
 1 2022-09-08 15:11:58 sp_needs  outside     18  39.2
 2 2022-06-24 15:41:58 lactating outside     25  39.1
 3 2022-06-24 15:41:58 sp_needs  outside     25  39.1
 4 2022-08-15 14:41:58 lactating outside     18  39.1
 5 2022-08-15 14:41:58 sp_needs  outside     18  39.1
 6 2022-09-08 15:11:58 lactating outside     18  39.1
 7 2022-06-24 15:56:58 sp_needs  outside     24  39
 8 2022-06-24 15:56:58 lactating outside     24  38.9
 9 2022-09-08 15:41:58 sp_needs  outside     18  38.9
10 2022-09-08 15:56:58 lactating outside     18  38.9
```

Your turn!

Order by **temp** and use **rh** as tie breaker. What was the lowest temperature?

Your turn!

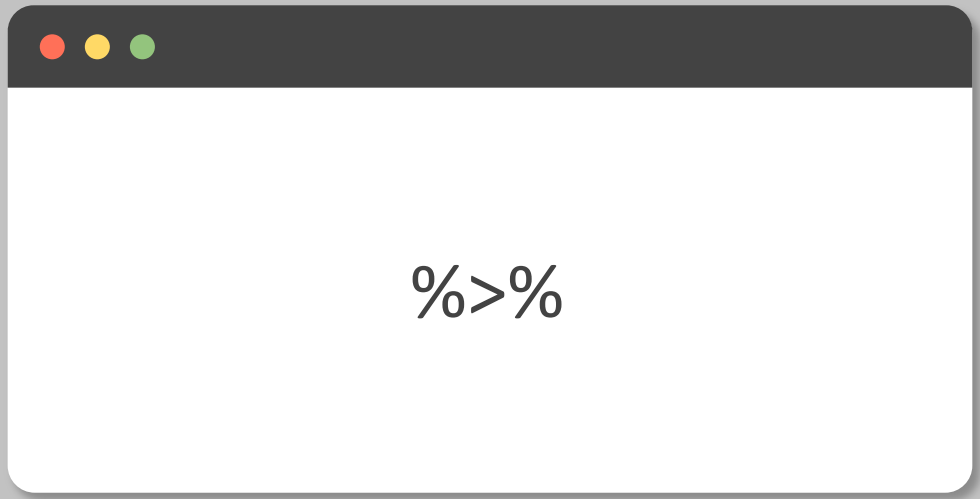
Order by **temp** and use **rh** as tie breaker. What was the lowest temperature?

arrange(env_data, temp)

```
# A tibble: 124,744 × 6
   date       time     barn      location    rh  temp
   <date>     <time>   <chr>     <chr>    <dbl> <dbl>
 1 2022-01-21 08:11:58 sp_needs  outside     77 -24
 2 2022-01-21 08:11:58 lactating outside     77 -23.9
 3 2022-01-29 05:41:58 sp_needs  outside     71 -23.9
 4 2022-01-29 05:41:58 lactating outside     71 -23.8
 5 2022-01-29 07:26:58 sp_needs  outside     72 -23.8
 6 2022-01-29 07:56:58 sp_needs  outside     73 -23.8
 7 2022-01-29 07:56:58 lactating outside     73 -23.7
 8 2022-01-21 05:41:58 lactating outside     76 -23.6
 9 2022-01-21 05:41:58 sp_needs  outside     77 -23.6
10 2022-01-29 07:26:58 lactating outside     72 -23.6
```

arrange(env_data, temp, rh)

```
# A tibble: 124,744 × 6
   date       time     barn      location    rh  temp
   <date>     <time>   <chr>     <chr>    <dbl> <dbl>
 1 2022-01-21 08:11:58 sp_needs  outside     77 -24
 2 2022-01-29 05:41:58 sp_needs  outside     71 -23.9
 3 2022-01-21 08:11:58 lactating outside     77 -23.9
 4 2022-01-29 05:41:58 lactating outside     71 -23.8
 5 2022-01-29 07:26:58 sp_needs  outside     72 -23.8
 6 2022-01-29 07:56:58 sp_needs  outside     73 -23.8
 7 2022-01-29 07:56:58 lactating outside     73 -23.7
 8 2022-01-29 07:26:58 lactating outside     72 -23.6
 9 2022-01-29 07:41:58 sp_needs  outside     73 -23.6
10 2022-01-21 05:41:58 lactating outside     76 -23.6
```



%>%

Use filter(), select() and arrange() to show only relative humidity from inside the Special Needs barn, ordered ascending by relative humidity

```
sp_needs <- filter(env_data, barn == "sp_needs", location == "inside")
sp_needs <- select(sp_needs , date, time, rh)
sp_needs <- arrange(sp_needs, rh)
sp_needs
```

```
# A tibble: 31,186 × 3
   date       time        rh
   <date>     <time>   <dbl>
 1 2022-05-12 13:11:58    29
 2 2022-05-12 15:26:58    29
 3 2022-05-12 16:11:58    29
 4 2022-05-12 12:56:58    30
 5 2022-05-12 13:26:58    30
 6 2022-05-12 13:41:58    30
 7 2022-05-12 13:56:58    30
 8 2022-05-12 14:11:58    30
 9 2022-05-12 14:26:58    30
10 2022-05-12 15:56:58    30
```

%>%

Use filter(), select() and arrange() to show only relative humidity from inside the Special Needs barn, ordered ascending by relative humidity

```
sp_needs <- filter(env_data, barn == "sp_needs", location == "inside")
sp_needs <- select(sp_needs , date, time, rh)
sp_needs <- arrange(sp_needs, rh)
sp_needs
```

```
# A tibble: 31,186 × 3
   date       time        rh
   <date>     <time>   <dbl>
 1 2022-05-12 13:11:58    29
 2 2022-05-12 15:26:58    29
 3 2022-05-12 16:11:58    29
 4 2022-05-12 12:56:58    30
 5 2022-05-12 13:26:58    30
 6 2022-05-12 13:41:58    30
 7 2022-05-12 13:56:58    30
 8 2022-05-12 14:11:58    30
 9 2022-05-12 14:26:58    30
10 2022-05-12 15:56:58    30
```

%>%

Use filter(), select() and arrange() to show only relative humidity from inside the Special Needs barn, ordered ascending by relative humidity

```
sp_needs <- filter(env_data, barn == "sp_needs", location == "inside")
sp_needs <- select(sp_needs , date, time, rh)
sp_needs <- arrange(sp_needs, rh)

sp_needs <- arrange(select(filter(env_data, barn == "sp_needs", location == "inside"), date, time, rh), rh)
```

%>%

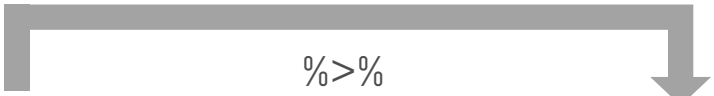
Use filter(), select() and arrange() to show only relative humidity from inside the Special Needs barn, ordered ascending by relative humidity

```
sp_needs <- filter(env_data, barn == "sp_needs", location == "inside")
sp_needs <- select(sp_needs , date, time, rh)
sp_needs <- arrange(sp_needs, rh)

sp_needs <- arrange(select(filter(env_data, barn == "sp_needs", location == "inside"), date, time, rh), rh)
```

%>%

The pipe operator %>% passes result on the left into first argument of function on the right.



env_data filter(_____, barn == "sp_needs", location == "inside")

```
filter(env_data, barn == "sp_needs", location == "inside")
```

equals to

```
env_data %>% filter(barn == "sp_needs", location == "inside")
```

Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

Your turn!

Use filter(), select(), arrange() and the pipe operator %>% to show only relative humidity from inside the Special Needs barn, ordered ascending by relative humidity

Your turn!

Use filter(), select(), arrange() and the pipe operator %>% to show only relative humidity from inside the Special Needs barn, ordered ascending by relative humidity

```
env_data %>%
  filter(barn == "sp_needs", location == "inside") %>%
  select(date, time, rh) %>%
  arrange(rh)
```

```
# A tibble: 31,186 × 3
   date       time        rh
   <date>     <time>   <dbl>
 1 2022-05-12 13:11:58    29
 2 2022-05-12 15:26:58    29
 3 2022-05-12 16:11:58    29
 4 2022-05-12 12:56:58    30
 5 2022-05-12 13:26:58    30
 6 2022-05-12 13:41:58    30
 7 2022-05-12 13:56:58    30
 8 2022-05-12 14:11:58    30
 9 2022-05-12 14:26:58    30
10 2022-05-12 15:56:58    30
```

Ultimate question

Are the temperatures inside the barns milder than outside?

i.e., warmer in the winter and colder in the summer

Ultimate question

Are the temperatures inside the barns milder than outside?

What do we need to know?

- Average temperatures during winter and summer months for each barn, inside and outside

barn	season	location	avg_temp
sp_needs	winter	inside	
sp_needs	winter	outside	
sp_needs	summer	inside	
sp_needs	summer	outside	
lactating	winter	inside	
lactating	winter	outside	
lactating	summer	inside	
lactating	summer	outside	

Ultimate question

Are the temperatures inside the barns milder than outside?

What do we need to know?

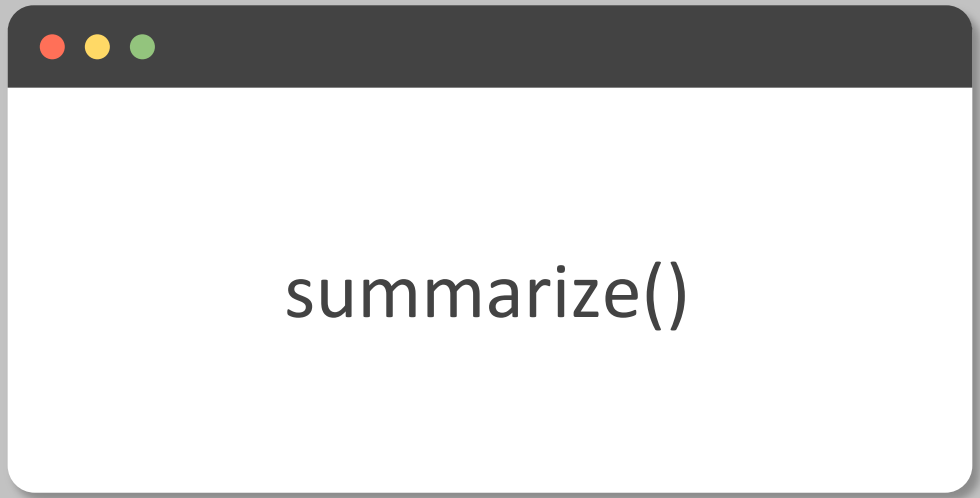
- Average temperatures during winter and summer months for each barn, inside and outside

barn	season	location	avg_temp
sp_needs	winter	inside	
sp_needs	winter	outside	
sp_needs	summer	inside	
sp_needs	summer	outside	
lactating	winter	inside	
lactating	winter	outside	
lactating	summer	inside	
lactating	summer	outside	

Deriving information

We can make table of summaries with **summarize()/summarise()**

We can make new variables with **mutate()**


summarize()

summarize()

Transforms a vector of data into one value

summarize(.data, new_column = function(vector))

.data	dataframe to transform
new_column	New column created by function()
function(...)	Function used to transform a vector
vector	Vector to be transformed, it's a column from .data

summarize()

Use summarize() to create a summary table with average and max temperatures

```
env_data %>%
  summarize(avg_temp = mean(temp),
                    max_temp = max(temp))
```

summarize()

Use summarize() to create a summary table with average and max temperatures

```
env_data %>%
  summarize(avg_temp = mean(temp),
                    max_temp = max(temp))
```

```
# A tibble: 1 × 2
  avg_temp max_temp
     <dbl>    <dbl>
1       NA       NA
```

summarize()

Use summarize() to create a summary table with average and max temperatures

```
env_data %>%
  summarize(avg_temp = mean(temp),
                    max_temp = max(temp))
```

```
# A tibble: 1 × 2
  avg_temp max_temp
     <dbl>    <dbl>
1       NA       NA
```

```
env_data %>% filter(is.na(temp))
```

```
# A tibble: 8 × 6
  date       time     barn      location    rh  temp
  <date>     <time>   <chr>     <chr>    <dbl> <dbl>
1 2022-01-31 10:41:58 lactating inside       NA    NA
2 2022-01-31 10:41:58 lactating outside      NA    NA
3 2022-01-31 10:41:58 sp_needs  inside       NA    NA
4 2022-01-31 10:41:58 sp_needs  outside      NA    NA
5 2022-01-31 10:56:58 lactating inside       NA    NA
6 2022-01-31 10:56:58 lactating outside      NA    NA
7 2022-01-31 10:56:58 sp_needs  inside       NA    NA
8 2022-01-31 10:56:58 sp_needs  outside      NA    NA
```

summarize()

Use summarize() to create a summary table with average and max temperatures

```
env_data %>%
  summarize(avg_temp = mean(temp),
            max_temp = max(temp))
```

```
# A tibble: 1 × 2
  avg_temp max_temp
     <dbl>    <dbl>
1       NA       NA
```

```
env_data %>%
  summarize(avg_temp = mean(temp, na.rm = TRUE),
            max_temp = max(temp, na.rm = TRUE))
```

OR

```
env_data %>%
  na.omit() %>%
  summarize(avg_temp = mean(temp),
            max_temp = max(temp))
```

```
# A tibble: 1 × 2
  avg_temp max_temp
     <dbl>    <dbl>
1     12.2     39.2
```

Your turn!

Using summarize() and filter(), get the min, mean, and max temperatures and relative humidity inside the Special Needs barn.
* Don't forget to account for NAs

Your turn!

Using summarize() and filter(), get the min, mean, and max temperatures and relative humidity inside the Special Needs barn.
* Don't forget to account for NAs

```
env_data %>%
  na.omit() %>%
  filter(barn == "sp_needs",
         location == "inside") %>%
  summarize(min_temp = min(temp),
            avg_temp = mean(temp),
            max_temp = max(temp),
            min_rh = min(rh),
            avg_rh = mean(rh),
            max_rh = max(rh))
```

```
# A tibble: 1 x 6
  min_temp avg_temp max_temp min_rh avg_rh max_rh
     <dbl>    <dbl>    <dbl>  <dbl>  <dbl>  <dbl>
1    -1.95     14.4     31.4     29   73.3     91
```

Your turn!

Using summarize() and filter(), get the min, mean, and max temperatures and relative humidity inside the Special Needs barn.
* Don't forget to account for NAs

```
env_data %>%
  na.omit() %>%
  filter(barn == "sp_needs",
         location == "inside") %>%
  summarize(min_temp = min(temp),
            avg_temp = mean(temp),
            max_temp = max(temp),
            min_rh = min(rh),
            avg_rh = mean(rh),
            max_rh = max(rh))
```

```
# A tibble: 1 x 6
  min_temp avg_temp max_temp min_rh avg_rh max_rh
     <dbl>    <dbl>    <dbl>  <dbl>  <dbl>  <dbl>
1    -1.95     14.4     31.4     29   73.3     91
```

Ultimate question

Are the temperatures inside the barns milder than outside?

What do we need to know?

- Average temperatures during winter and summer months for each barn, inside and outside

barn	season	location	avg_temp
sp_needs	winter	inside	
sp_needs	winter	outside	
sp_needs	summer	inside	
sp_needs	summer	outside	
lactating	winter	inside	
lactating	winter	outside	
lactating	summer	inside	
lactating	summer	outside	


group_by()

summarize()

Group observations by common values

summarize(.data, ...)

.data	dataframe to transform
...	one or more column names to group

group_by()

Get the min, mean, and max temperatures per barn and location

```
env_data %>%
  na.omit() %>%
  summarize(min_temp = min(temp),
            avg_temp = mean(temp),
            max_temp = max(temp
```

```
# A tibble: 1 × 3
  min_temp avg_temp max_temp
     <dbl>    <dbl>    <dbl>
1      -24     12.2     39.2
```

group_by()

Get the min, mean, and max temperatures per barn and location

```
env_data %>%
  na.omit() %>%
  group_by(barn, location) %>%
  summarize(min_temp = min(temp),
            avg_temp = mean(temp),
            max_temp = max(temp
```

```
# A tibble: 4 x 5
# Groups:   barn [2]
  barn      location min_temp avg_temp max_temp
  <chr>     <chr>       <dbl>    <dbl>    <dbl>
1 lactating inside       1.34     13.8     31.3
2 lactating outside    -23.9      10.3     39.1
3 sp_needs  inside      -1.95     14.4     31.4
4 sp_needs  outside     -24        10.3     39.2
```

group_by()

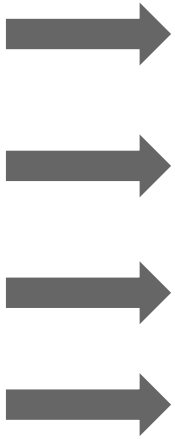
```
sampled_env_data %>% na.omit() %>% summarize(min = min(temp), avg = mean(temp), max = max(temp))
```

barn	location	rh	temp
lactating	inside	79	20.4
lactating	inside	85	19.9
lactating	outside	83	-2.5
lactating	outside	47	23.4
sp_needs	inside	83	12.0
sp_needs	inside	78	9.2
sp_needs	outside	73	-14.5
sp_needs	outside	58	15.7

min	avg	max
-14.5	10.4	23.4

group_by() + summarize()

barn	location	rh	temp
lactating	inside	79	20.4
lactating	inside	85	19.9
lactating	outside	83	-2.5
lactating	outside	47	23.4
sp_needs	inside	83	12.0
sp_needs	inside	78	9.2
sp_needs	outside	73	-14.5
sp_needs	outside	58	15.7



min	avg	max
19.9	20.1	20.4
-2.5	10.4	23.4
9.2	10.6	12.0
-14.5	0.6	15.7

Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

group_by()

```
sampled_env_data %>% group_by(barn, location) %>% na.omit() %>%
        summarize(min = min(temp),
                  avg = mean(temp),
                  max = max(temp))
```

barn	location	rh	temp
lactating	inside	79	20.4
lactating	inside	85	19.9
lactating	outside	83	-2.5
lactating	outside	47	23.4
sp_needs	inside	83	12.0
sp_needs	inside	78	9.2
sp_needs	outside	73	-14.5
sp_needs	outside	58	15.7

min	avg	max
19.9	20.1	20.4
-2.5	10.4	23.4
9.2	10.6	12.0
-14.5	0.6	15.7

barn	location	min	avg	max
lactating	inside	19.9	20.1	20.4
lactating	outside	-2.5	10.4	23.4
sp_needs	inside	9.2	10.6	12.0
sp_needs	outside	-14.5	0.6	15.7

Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

Your turn!

Use group_by(), filter(), and summarize() to show the lowest and highest relative humidity and temperature of the inside of each barn

Your turn!

Use group_by(), filter(), and summarize() to show the lowest and highest relative humidity and temperature of the inside of each barn

```
env_data %>%
  filter(location == "inside") %>%
  na.omit() %>%
  group_by(barn, location) %>%
  summarize(min_temp = min(temp),
            max_temp = max(temp),
            min_rh = min(rh),
            max_rh = max(rh))
```

```
# A tibble: 2 × 6
# Groups:   barn [2]
  barn      location min_temp max_temp min_rh max_rh
  <chr>     <chr>       <dbl>    <dbl>  <dbl>  <dbl>
1 lactating inside       1.34     31.3     24     90
2 sp_needs  inside      -1.95     31.4     29     91
```


mutate()

mutate()

Apply vectorized functions to columns to create new columns

mutate(.data, new_column = function(vector))

.data	dataframe to transform
new_column	New column created by function()
function(...)	Function used to transform a vector
vector	Vector to be transformed, can be a column from .data

mutate()

Create new columns

```
env_data %>%
  mutate(year = lubridate::year(date),
         month = lubridate::month(date),
         day = lubridate::day(date),
         barn = dplyr::if_else(barn == "sp_needs", "special_needs", barn))
```

```
# A tibble: 124,744 × 9
   date       time    barn          location    rh  temp  year month   day
   <date>     <time>  <chr>         <chr>    <dbl> <dbl> <dbl> <dbl> <int>
 1 2022-01-01 11'58" lactating     inside      74  7.74  2022     1     1
 2 2022-01-01 11'58" lactating     outside     87  2.5   2022     1     1
 3 2022-01-01 11'58" special_needs inside      78 10.1   2022     1     1
 4 2022-01-01 11'58" special_needs outside     87  2.5   2022     1     1
 5 2022-01-01 26'58" lactating     inside      74  8.31  2022     1     1
```

Adapted from: https://tidy-ds.wjakethompson.com and https://r4ds.had.co.nz

Recap!

Extract variables with **select()**

Extract observations with **filter()**

Arrange/Sort observations with **arrange()**

Make table of summaries with **summarize**()

Make new variables with **mutate()**

Your turn!

Using the functions from today, create the following table to answer our ultimate question:
Are the temperatures inside the barns milder than outside?

barn	season	location	avg_temp
sp_needs	winter	inside	
sp_needs	winter	outside	
sp_needs	summer	inside	
sp_needs	summer	outside	
lactating	winter	inside	
lactating	winter	outside	
lactating	summer	inside	
lactating	summer	outside	

Your turn!

Using the functions from today, create the following table to answer our ultimate question:
Are the temperatures inside the barns milder than outside?

```
env_data %>%
  na.omit() %>%
  mutate(season = if_else(date >= "2021-12-31" & date <= "2022-03-20",
                          true = "winter",
                          false = if_else(date >= "2022-06-21" & date <= "2022-09-23",
                                          true = "summer",
                                          false = "spring/fall"))) %>%
  filter(season %in% c("summer", "winter")) %>%
  group_by(barn, season, location) %>%
  summarize(avg_temp = mean(temp)) %>%
  arrange(desc(barn), desc(season))
```

Your turn!

Using the functions from today, create the following table to answer our ultimate question:
Are the temperatures inside the barns milder than outside?

```
# A tibble: 8 × 4
# Groups:   barn, season [4]
  barn      season location avg_temp
  <chr>     <chr>  <chr>       <dbl>
1 sp_needs  winter inside       7.87
2 sp_needs  winter outside     -5.90
3 sp_needs  summer inside      20.5
4 sp_needs  summer outside     21.9
5 lactating winter inside       7.62
6 lactating winter outside     -5.88
7 lactating summer inside      20.3
8 lactating summer outside     21.8
```

Your turn!

Using the functions from today, create the following table to answer our ultimate question:
Are the temperatures inside the barns milder than outside?

```
# A tibble: 8 × 4
# Groups:   barn, season [4]
  barn      season location avg_temp
  <chr>     <chr>  <chr>       <dbl>
1 sp_needs  winter inside       7.87
2 sp_needs  winter outside     -5.90
3 sp_needs  summer inside      20.5
4 sp_needs  summer outside     21.9
5 lactating winter inside       7.62
6 lactating winter outside     -5.88
7 lactating summer inside      20.3
8 lactating summer outside     21.8
```

Yes!*

*I'll leave it up to your curiosity to check the statistical significance ☺

Coffee Break!