R workshop

Jackie Maud

21 January 2019

# Day 2

## Tidy Data in spreadsheets

The functions for tidying data are:

:: tells which package to use function from

*tidyr::spread()* tidyr::gather()

### Spread

3 principal arguments

1. the data
2. the *key* column variable will become the new column names
3. the *value* column variable whch will fill the new column variables

Use surveys dataset

Make from LONG to WIDE (spread)

surveys<-read\_csv(here::here("read\_data", "surveys.csv"))

## Parsed with column specification:  
## cols(  
## record\_id = col\_double(),  
## month = col\_double(),  
## day = col\_double(),  
## year = col\_double(),  
## plot\_id = col\_double(),  
## species\_id = col\_character(),  
## sex = col\_character(),  
## hindfoot\_length = col\_double(),  
## weight = col\_double()  
## )

library(tidyverse)  
  
#create a wide data format of surveys using spread  
  
#first create a summary  
  
surveys\_gw <- surveys %>%  
 drop\_na(weight) %>%   
 group\_by(species\_id) %>%   
 summarize(mean\_weight = mean(weight))  
  
  
str(surveys\_gw)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 25 obs. of 2 variables:  
## $ species\_id : chr "BA" "DM" "DO" "DS" ...  
## $ mean\_weight: num 8.6 43.2 48.9 120.1 159.2 ...

wide\_surveys\_gw <-surveys\_gw %>%   
 spread(key = species\_id, value = mean\_weight)

### gather()

# Now back to long data from wide

gather takes 4 arguments

1. data
2. *key*
3. *value*
4. names of columns we use to fill the key variable (or drop)

long\_surveys\_gw <-wide\_surveys\_gw %>%   
 gather(key = species\_id, value = mean\_weight)

## Sending Tidy Data

### Changelog

* Update your change log with changes to raw data/project

### Data dictionary

\*Create to define our variables

tidy\_gsi <- read\_csv(here::here("read\_data", "tidy\_gsi.csv"))

## Parsed with column specification:  
## cols(  
## hakai\_id = col\_character(),  
## stock\_1 = col\_character(),  
## region\_1 = col\_double(),  
## prob\_1 = col\_double(),  
## stock\_2 = col\_character(),  
## region\_2 = col\_double(),  
## prob\_2 = col\_double(),  
## stock\_3 = col\_character(),  
## region\_3 = col\_double(),  
## prob\_3 = col\_double(),  
## stock\_4 = col\_character(),  
## region\_4 = col\_double(),  
## prob\_4 = col\_double(),  
## stock\_5 = col\_character(),  
## region\_5 = col\_double(),  
## prob\_5 = col\_double()  
## )

view(tidy\_gsi)

# Analysing data

## Importing from Hakai Data Portal

Switched to data\_wrangling script to import data into our read\_data file

Chl\_a, fish and sockeye stock ID data

library(here)

## here() starts at C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2

fish <-read\_csv(here("read\_data", "fish.csv"))

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## X1 = col\_double(),  
## action = col\_logical(),  
## date = col\_date(format = ""),  
## package\_id = col\_logical(),  
## fish\_time\_out = col\_logical(),  
## fish\_time\_dewar = col\_logical(),  
## fork\_length\_field = col\_double(),  
## height\_field = col\_double(),  
## weight\_field = col\_logical(),  
## date\_processed = col\_date(format = ""),  
## weight = col\_double(),  
## standard\_length = col\_double(),  
## fork\_length = col\_double(),  
## photo\_number = col\_logical(),  
## comments = col\_logical(),  
## quality\_log = col\_logical()  
## )

## See spec(...) for full column specifications.

## Warning: 2861 parsing failures.  
## row col expected actual file  
## 1393 photo\_number 1/0/T/F/TRUE/FALSE 3142 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/fish.csv'  
## 2093 photo\_number 1/0/T/F/TRUE/FALSE 2835 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/fish.csv'  
## 2247 photo\_number 1/0/T/F/TRUE/FALSE 3204 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/fish.csv'  
## 2527 photo\_number 1/0/T/F/TRUE/FALSE 3137 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/fish.csv'  
## 2745 photo\_number 1/0/T/F/TRUE/FALSE 2009 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/fish.csv'  
## .... ............ .................. ...... ..........................................................................  
## See problems(...) for more details.

chla <- read\_csv(here("read\_data", "chla.csv"))

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## action = col\_logical(),  
## date = col\_date(format = ""),  
## work\_area = col\_character(),  
## survey = col\_character(),  
## site\_id = col\_character(),  
## gather\_lat = col\_logical(),  
## gather\_long = col\_logical(),  
## collection\_method = col\_logical(),  
## pressure\_transducer\_depth = col\_logical(),  
## collected = col\_datetime(format = ""),  
## preserved = col\_datetime(format = ""),  
## analyzed = col\_datetime(format = ""),  
## lab\_technician = col\_character(),  
## project\_specific\_id = col\_character(),  
## hakai\_id = col\_character(),  
## is\_blank = col\_logical(),  
## is\_solid\_standard = col\_logical(),  
## filter\_size\_mm = col\_logical(),  
## filter\_type = col\_character(),  
## calibration = col\_datetime(format = "")  
## # ... with 8 more columns  
## )  
## See spec(...) for full column specifications.

## Warning: 15289 parsing failures.  
## row col expected actual file  
## 2627 gather\_lat 1/0/T/F/TRUE/FALSE 50.11505 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/chla.csv'  
## 2627 gather\_long 1/0/T/F/TRUE/FALSE -125.22168 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/chla.csv'  
## 2628 gather\_lat 1/0/T/F/TRUE/FALSE 50.11505 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/chla.csv'  
## 2628 gather\_long 1/0/T/F/TRUE/FALSE -125.22168 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/chla.csv'  
## 2629 gather\_lat 1/0/T/F/TRUE/FALSE 50.11505 'C:/Users/HP/Documents/R/R projects\_Jackie/R-workshop2/read\_data/chla.csv'  
## .... ........... .................. .......... ..........................................................................  
## See problems(...) for more details.

tidy\_gsi <- read\_csv(here("read\_data", "tidy\_gsi.csv"))

## Parsed with column specification:  
## cols(  
## hakai\_id = col\_character(),  
## stock\_1 = col\_character(),  
## region\_1 = col\_double(),  
## prob\_1 = col\_double(),  
## stock\_2 = col\_character(),  
## region\_2 = col\_double(),  
## prob\_2 = col\_double(),  
## stock\_3 = col\_character(),  
## region\_3 = col\_double(),  
## prob\_3 = col\_double(),  
## stock\_4 = col\_character(),  
## region\_4 = col\_double(),  
## prob\_4 = col\_double(),  
## stock\_5 = col\_character(),  
## region\_5 = col\_double(),  
## prob\_5 = col\_double()  
## )

fish %>%   
 count(species)

## # A tibble: 6 x 2  
## species n  
## <chr> <int>  
## 1 CK 12  
## 2 CO 98  
## 3 CU 1689  
## 4 HE 282  
## 5 PI 860  
## 6 SO 3497

fish\_d09 <- fish %>%   
 filter(site\_id == "D09") %>%   
 select(hakai\_id, jsp\_survey\_id, seine\_id, date, species, site\_id, fork\_length, weight) %>%   
 mutate(k = (10^5 \* weight) / fork\_length^3) %>%   
 drop\_na(k)

### Annoying things that will get you

## Factors

str(fish\_d09)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 832 obs. of 9 variables:  
## $ hakai\_id : chr "U4802" "U4776" "U4728" "U4801" ...  
## $ jsp\_survey\_id: chr "DE112" "DE112" "DE112" "DE112" ...  
## $ seine\_id : chr "DE112N1" "DE112N1" "DE112N1" "DE112N1" ...  
## $ date : Date, format: "2015-05-20" "2015-05-20" ...  
## $ species : chr "SO" "SO" "SO" "SO" ...  
## $ site\_id : chr "D09" "D09" "D09" "D09" ...  
## $ fork\_length : num 106 106 97 102 102 97 96 95 128 101 ...  
## $ weight : num 10.1 11.3 8.8 9.9 8.7 8.4 7.9 8.1 19 9.8 ...  
## $ k : num 0.848 0.949 0.964 0.933 0.82 ...

#or  
  
class(fish\_d09$species)

## [1] "character"

#coerce a column to be a factor  
  
fish\_d09$species <- factor(fish\_d09$species)  
  
levels(fish\_d09$species)

## [1] "CO" "CU" "HE" "PI" "SO"

If you have factors that are numbers, don’t try to do maths with these Under the hood R will treat your factor levels as numbers.

## Dates

read\_csv(): treats ISO date standards (yyyy-mm-dd) as a DATE object read.csv(): treats them as characters - not ideal

Lubridate package - to help with dates

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:here':  
##   
## here

## The following object is masked from 'package:base':  
##   
## date

# Extract data components (day, month, year, Julian day/yday)  
  
fish\_d09 <- fish\_d09 %>%   
 mutate(year = year(date),  
 month = month(date),  
 week = week(date),  
 yday = yday(date))  
  
# to change format of column to DATE  
  
#fish\_d09$date <- as.date(fish\_d09$date)

# Can do maths with lubridate

* periods *intervals* durations

## Joining data

Data:

*chla* tidy\_gsi \*fish\_d09

# see dplyr cheatsheet for help on this  
  
left\_join(fish\_d09, tidy\_gsi, by = "hakai\_id")

## # A tibble: 832 x 28  
## hakai\_id jsp\_survey\_id seine\_id date species site\_id fork\_length  
## <chr> <chr> <chr> <date> <fct> <chr> <dbl>  
## 1 U4802 DE112 DE112N1 2015-05-20 SO D09 106  
## 2 U4776 DE112 DE112N1 2015-05-20 SO D09 106  
## 3 U4728 DE112 DE112N1 2015-05-20 SO D09 97  
## 4 U4801 DE112 DE112N1 2015-05-20 SO D09 102  
## 5 U4777 DE112 DE112N1 2015-05-20 SO D09 102  
## 6 U4779 DE112 DE112N1 2015-05-20 SO D09 97  
## 7 U4778 DE112 DE112N1 2015-05-20 SO D09 96  
## 8 U4800 DE112 DE112N1 2015-05-20 SO D09 95  
## 9 U4780 DE112 DE112N1 2015-05-20 SO D09 128  
## 10 U350 DE112 DE112N1 2015-05-20 SO D09 101  
## # ... with 822 more rows, and 21 more variables: weight <dbl>, k <dbl>,  
## # year <dbl>, month <dbl>, week <dbl>, yday <dbl>, stock\_1 <chr>,  
## # region\_1 <dbl>, prob\_1 <dbl>, stock\_2 <chr>, region\_2 <dbl>,  
## # prob\_2 <dbl>, stock\_3 <chr>, region\_3 <dbl>, prob\_3 <dbl>,  
## # stock\_4 <chr>, region\_4 <dbl>, prob\_4 <dbl>, stock\_5 <chr>,  
## # region\_5 <dbl>, prob\_5 <dbl>

right\_join(fish\_d09, tidy\_gsi, by = "hakai\_id")

## # A tibble: 1,187 x 28  
## hakai\_id jsp\_survey\_id seine\_id date species site\_id fork\_length  
## <chr> <chr> <chr> <date> <fct> <chr> <dbl>  
## 1 U10 <NA> <NA> NA <NA> <NA> NA  
## 2 U16 <NA> <NA> NA <NA> <NA> NA  
## 3 U17 <NA> <NA> NA <NA> <NA> NA  
## 4 U21 <NA> <NA> NA <NA> <NA> NA  
## 5 U25 <NA> <NA> NA <NA> <NA> NA  
## 6 U31 <NA> <NA> NA <NA> <NA> NA  
## 7 U35 <NA> <NA> NA <NA> <NA> NA  
## 8 U42 <NA> <NA> NA <NA> <NA> NA  
## 9 U43 <NA> <NA> NA <NA> <NA> NA  
## 10 U7 <NA> <NA> NA <NA> <NA> NA  
## # ... with 1,177 more rows, and 21 more variables: weight <dbl>, k <dbl>,  
## # year <dbl>, month <dbl>, week <dbl>, yday <dbl>, stock\_1 <chr>,  
## # region\_1 <dbl>, prob\_1 <dbl>, stock\_2 <chr>, region\_2 <dbl>,  
## # prob\_2 <dbl>, stock\_3 <chr>, region\_3 <dbl>, prob\_3 <dbl>,  
## # stock\_4 <chr>, region\_4 <dbl>, prob\_4 <dbl>, stock\_5 <chr>,  
## # region\_5 <dbl>, prob\_5 <dbl>

inner\_join(fish\_d09, tidy\_gsi, by = "hakai\_id") #rows that have all required data

## # A tibble: 147 x 28  
## hakai\_id jsp\_survey\_id seine\_id date species site\_id fork\_length  
## <chr> <chr> <chr> <date> <fct> <chr> <dbl>  
## 1 U350 DE112 DE112N1 2015-05-20 SO D09 101  
## 2 U349 DE112 DE112N1 2015-05-20 SO D09 104  
## 3 U357 DE112 DE112N1 2015-05-20 SO D09 101  
## 4 U355 DE112 DE112N1 2015-05-20 SO D09 98  
## 5 U362 DE112 DE112N1 2015-05-20 SO D09 89  
## 6 U356 DE112 DE112N1 2015-05-20 SO D09 103  
## 7 U363 DE112 DE112N1 2015-05-20 SO D09 101  
## 8 U347 DE112 DE112N1 2015-05-20 SO D09 102  
## 9 U361 DE112 DE112N1 2015-05-20 SO D09 98  
## 10 U319 DE121 DE121N1 2015-05-24 SO D09 102  
## # ... with 137 more rows, and 21 more variables: weight <dbl>, k <dbl>,  
## # year <dbl>, month <dbl>, week <dbl>, yday <dbl>, stock\_1 <chr>,  
## # region\_1 <dbl>, prob\_1 <dbl>, stock\_2 <chr>, region\_2 <dbl>,  
## # prob\_2 <dbl>, stock\_3 <chr>, region\_3 <dbl>, prob\_3 <dbl>,  
## # stock\_4 <chr>, region\_4 <dbl>, prob\_4 <dbl>, stock\_5 <chr>,  
## # region\_5 <dbl>, prob\_5 <dbl>

anti\_join(fish\_d09, tidy\_gsi, by = "hakai\_id") #rows that DON'T have a match, i.e. NAs

## # A tibble: 685 x 13  
## hakai\_id jsp\_survey\_id seine\_id date species site\_id fork\_length  
## <chr> <chr> <chr> <date> <fct> <chr> <dbl>  
## 1 U4802 DE112 DE112N1 2015-05-20 SO D09 106  
## 2 U4776 DE112 DE112N1 2015-05-20 SO D09 106  
## 3 U4728 DE112 DE112N1 2015-05-20 SO D09 97  
## 4 U4801 DE112 DE112N1 2015-05-20 SO D09 102  
## 5 U4777 DE112 DE112N1 2015-05-20 SO D09 102  
## 6 U4779 DE112 DE112N1 2015-05-20 SO D09 97  
## 7 U4778 DE112 DE112N1 2015-05-20 SO D09 96  
## 8 U4800 DE112 DE112N1 2015-05-20 SO D09 95  
## 9 U4780 DE112 DE112N1 2015-05-20 SO D09 128  
## 10 U348 DE112 DE112N1 2015-05-20 SO D09 94  
## # ... with 675 more rows, and 6 more variables: weight <dbl>, k <dbl>,  
## # year <dbl>, month <dbl>, week <dbl>, yday <dbl>

#view() displays results for last function (if haven't created it as new df)

## ggplot2

To build a ggplot:

ggplot(data = DATA, mapping = aes(MAPPINGS)) + GEOM\_FUNCTION()

Example:

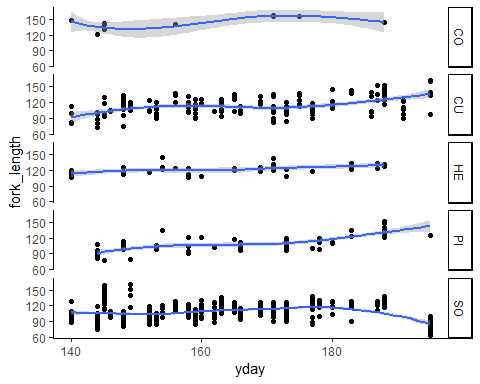
# ggplot(data = surveys, mapping = aes(species, weight)) +

#+geom\_point()

ggplot(fish\_d09, aes()) +  
 geom\_point(aes(x = yday, y = fork\_length))+  
 geom\_smooth(aes(x = yday, y = fork\_length), model = lm)+  
 theme\_classic()+  
 facet\_grid(species~.) #separates data by specified variable

## Warning: Ignoring unknown parameters: model

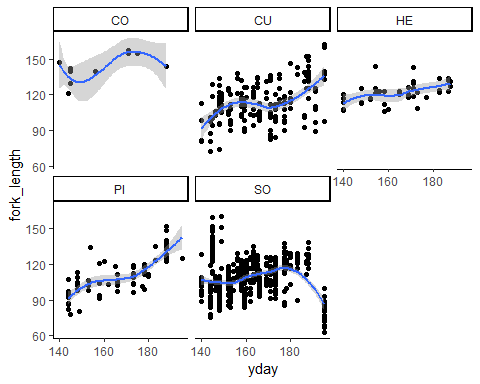
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



#or facet\_wrap(species~.)  
  
ggplot(fish\_d09, aes()) +  
 geom\_point(aes(x = yday, y = fork\_length))+  
 geom\_smooth(aes(x = yday, y = fork\_length), model = lm)+  
 theme\_classic()+  
 facet\_wrap(species~.)

## Warning: Ignoring unknown parameters: model

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



# check out different bg themes, \_bw (gridlines) and \_classic (no gridlines)

Cookbook for R <http://www.cookbook-r.com/>

How to adjust legends, axes, etc.

Geom list:

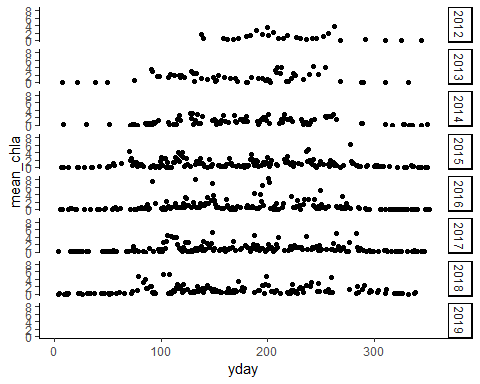
tidyverse.org/reference

<https://dplyr.tidyverse.org/reference/index.html>

## Playtime

chla\_date <- chla %>%   
 group\_by(date) %>%   
 summarize(mean\_chla = mean(chla, na.rm = TRUE))  
  
chla\_date <- chla\_date %>%   
 mutate(year = year(date),  
 month = month(date),  
 week = week(date),  
 yday = yday(date))  
   
  
ggplot(chla\_date, aes()) +  
 geom\_point(aes(x = yday, y = mean\_chla))+  
 theme\_classic()+  
 facet\_grid(year~.)

## Warning: Removed 51 rows containing missing values (geom\_point).



ggplot(chla\_date, aes()) +  
 geom\_point(aes(x = yday, y = mean\_chla))+  
 theme\_classic()+  
 facet\_wrap(year~.)

## Warning: Removed 51 rows containing missing values (geom\_point).

