RWorkshop_Tidying_Data

Keith Bouma-Gregson

1/14/2017

Tidy data

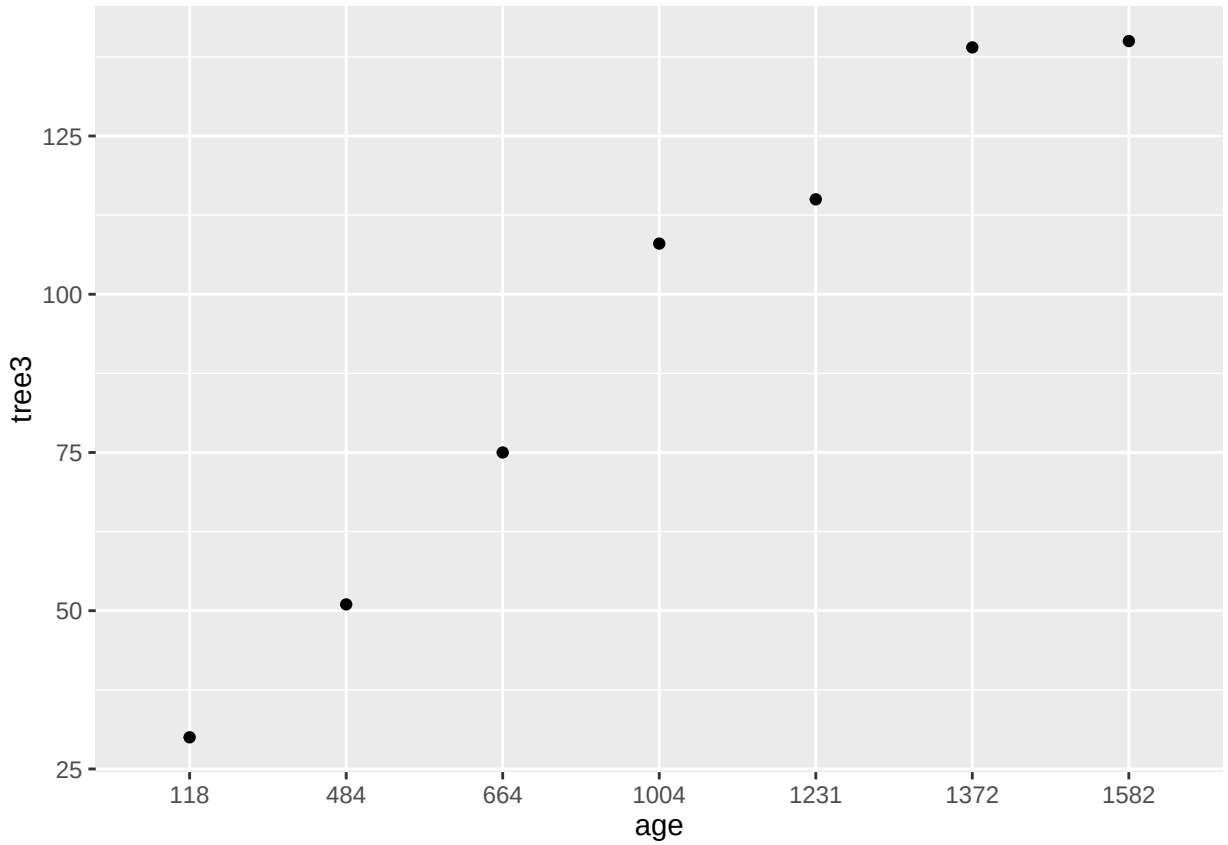
- Tidy data is
  - each column is a variable
  - each row is an observation

This is also known as **long** format vs. **wide** format where observations are spread among rows, not columns.

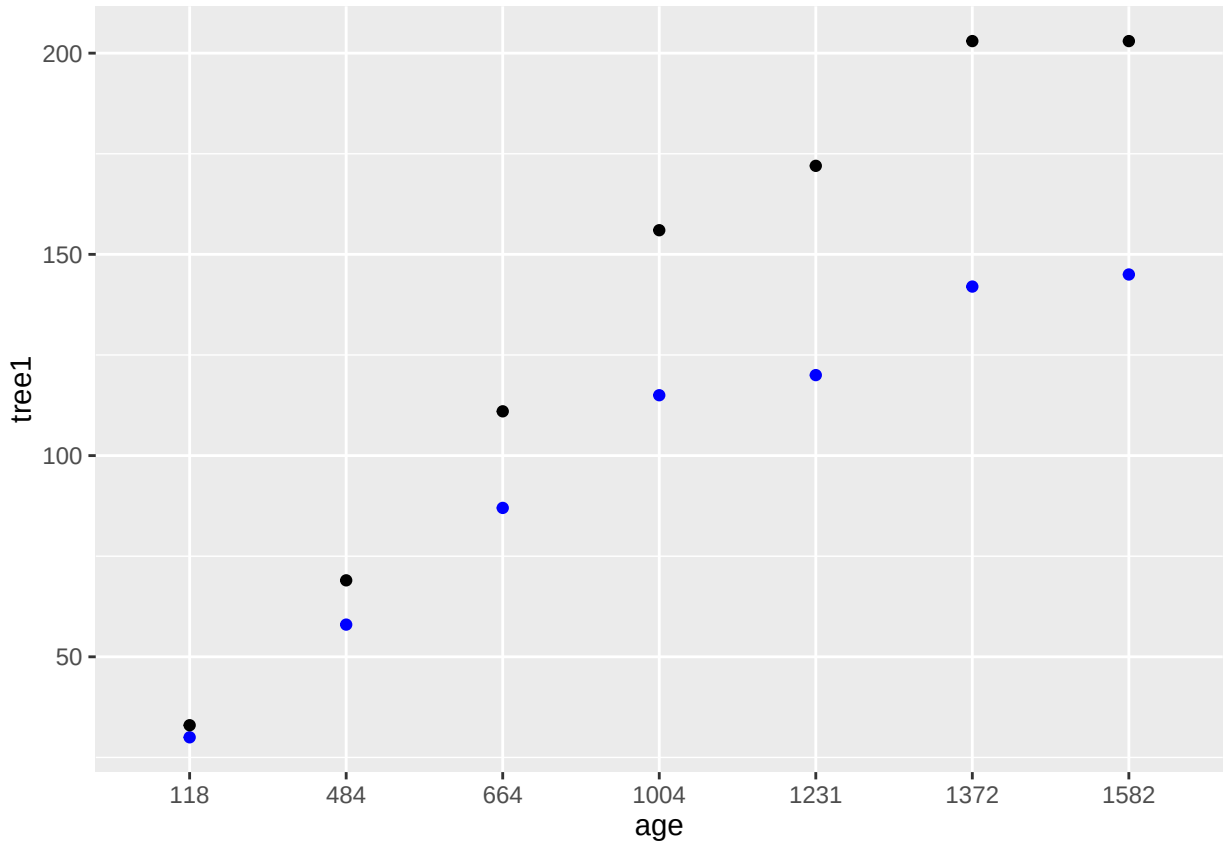
```
# Look at our data
head(Orange.messy)
```

```
##    age tree1 tree2 tree3 tree4 tree5
## 1  118    30    33    30    32    30
## 2  484    58    69    51    62    49
## 3  664    87   111    75   112    81
## 4 1004   115   156   108   167   125
## 5 1231   120   172   115   179   142
## 6 1372   142   203   139   209   174
```

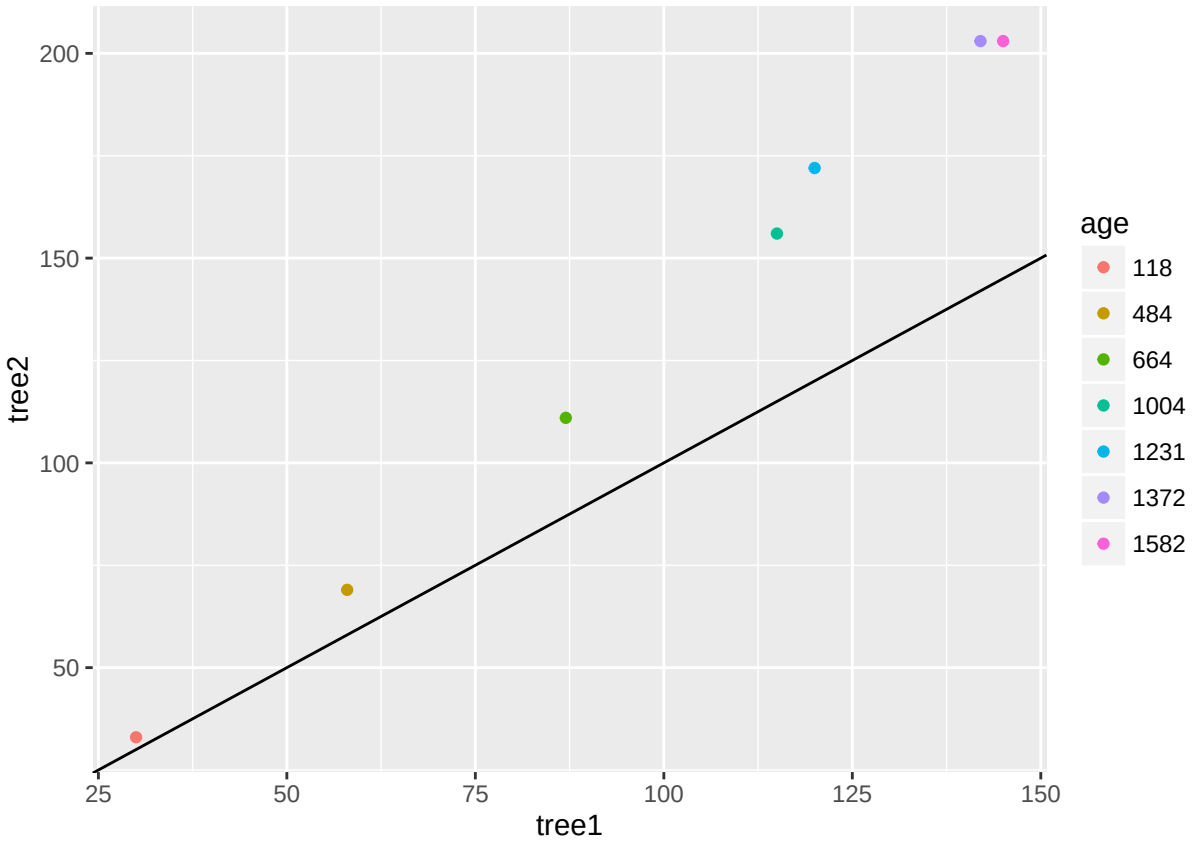
```
library(ggplot2)
# Plots with Orange.messy
ggplot(Orange.messy, aes(x= age, y= tree3)) + geom_point()
```




```
ggplot(data= Orange.messy) +
  geom_point(aes(x= age, y= tree1), color= "blue") +
  geom_point(aes(x= age, y= tree2), color= "black")
```




```
ggplot(data= Orange.messy, aes(x= tree1, y= tree2)) +
  geom_point(aes(color= age)) +
  geom_abline(intercept= 0, slope= 1)
```



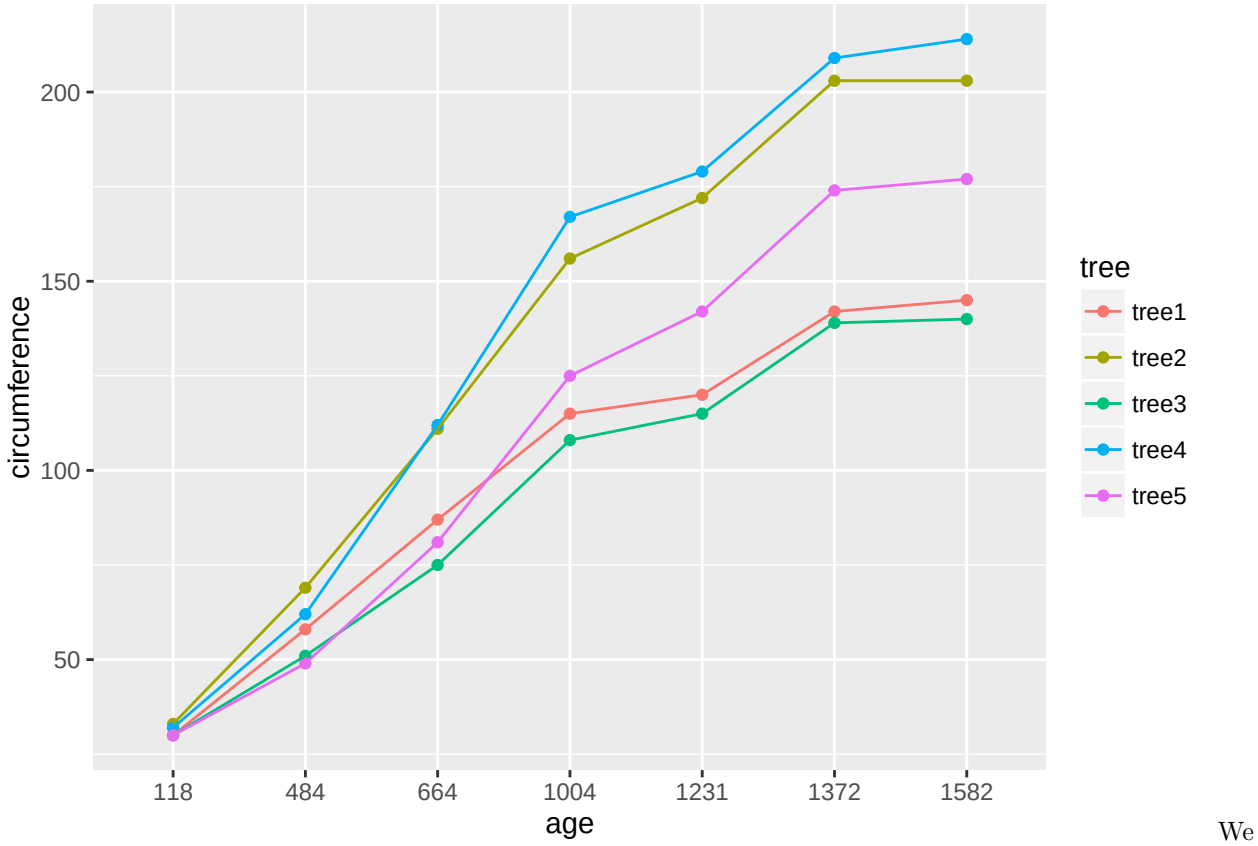
Tidying data

Enter the packages `tidyr()` and `reshape2()`. Which will move data between wide and long formats by specifiying the id, variable, and value columns.

```
## gather() function in tidyr
library(tidyr)
Orange.tidy1 <- gather(data= Orange.messy,
                       key= tree,
                       value= circumference,
                       tree1:tree5)

Orange.tidy2 <- gather(data= Orange.messy,
                       key= tree,
                       value= circumference,
                       -age)


## melt function in reshape2
library(reshape2)
Orange.melt1 <- melt(data= Orange.messy,
                     variable.name= "tree", # key
                     value.name= "circumference", # value
                     id.vars= "age")
```

Now plot again

```
ggplot(data= Orange.tidy1, aes(x= age, y=circumference, group= tree)) +
  geom_point(aes(color= tree)) +
  geom_line(aes(color= tree))
```



We can also move from long to wide with `spread`

```
orange.wide <- spread(data= Orange.tidy1, key= tree, value= circumference)
```

Exercise

Make the `messy` data frame into long format. This data shows the mass (g) of males and females from 2 treatments and 4 sites.

```
head(messy)
```

```
##   id site     trt   mass.M   mass.F
## 1  1    1 control 30.14956 22.29442
## 2  2    2 control 26.13537 17.30705
## 3  3    3 control 28.53815 20.50220
## 4  4    4 control 33.86204 12.11137
## 5  5    1 control 21.70272 13.11700
## 6  6    2 control 24.50873 19.13294
```

The long form should like something like this

```
Messy.tidy <- gather(data= messy, key= metric, value= value, mass.M:mass.F)
head(Messy.tidy)
```

```
##   id site     trt metric    value
## 1  1    1 control mass.M 30.14956
```

```
## 2  2    2 control mass.M 26.13537
## 3  3    3 control mass.M 28.53815
## 4  4    4 control mass.M 33.86204
## 5  5    1 control mass.M 21.70272
## 6  6    2 control mass.M 24.50873
```

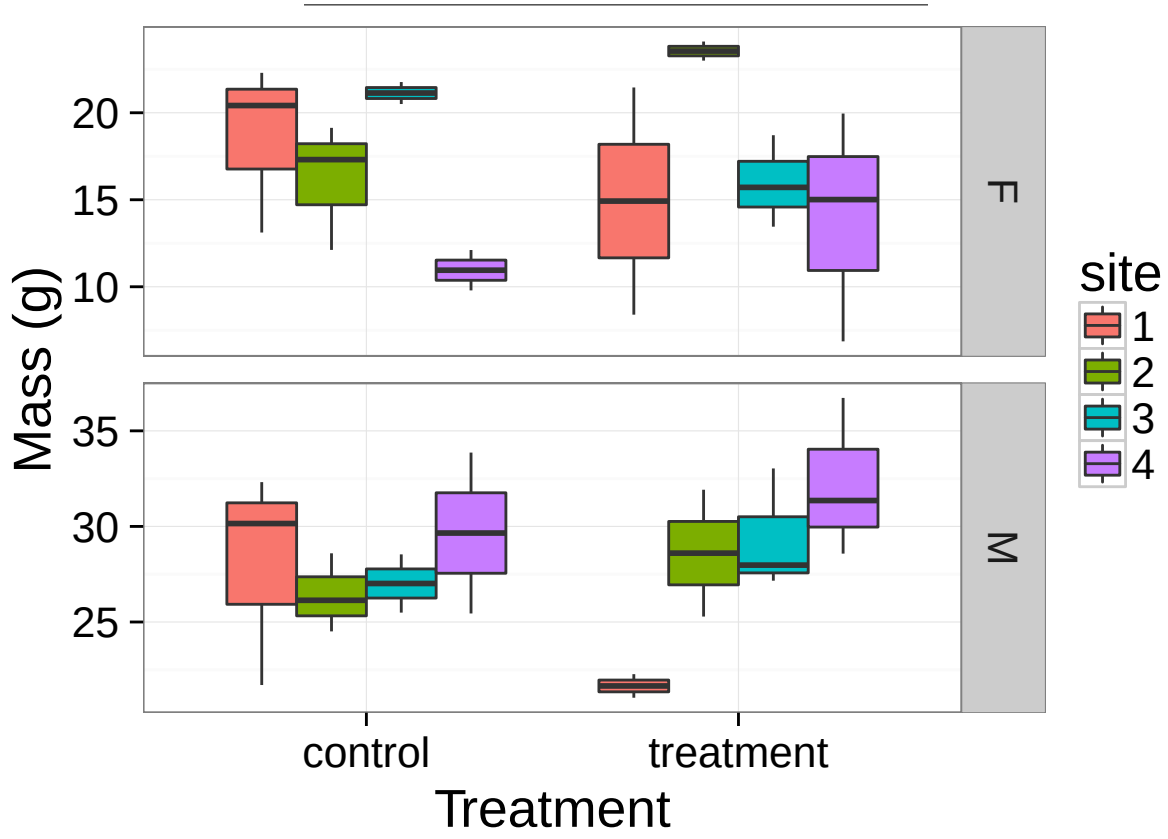
Separating data with the `separate` function. This will split a variable name at a particular character and turn it into two variable columns.

```
#
tidy.data <- separate(data= Messy.tidy, col= metric, into= c("metric", "sex"), sep="\\.")

tidy.data$site <- as.factor(tidy.data$site)

str(tidy.data)
```

```
## 'data.frame':    40 obs. of  6 variables:
##  $ id    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ site  : Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 4 1 2 ...
##  $ trt   : Factor w/ 2 levels "control","treatment": 1 1 1 1 1 1 1 1 1 1 ...
##  $ metric: chr  "mass" "mass" "mass" "mass" ...
##  $ sex   : chr  "M" "M" "M" "M" ...
##  $ value : num  30.1 26.1 28.5 33.9 21.7 ...
```

Exercise

Using the `tidy.data` data frame, make a ggplot boxplot with x= treatment, y= mass, with a separate color for each site and faceted by sex. Then make it more aesthetically pleasing than the default settings.



```
## Plot the data
ggplot(data= tidy.data, aes(x= trt, y= value)) +
  geom_boxplot(aes(fill= site)) +
  labs(x= "Treatment", y= "Mass (g)") +
  facet_grid(sex~., scales= "free_y") +
  theme_bw(base_size= 20)
```