Tidy data

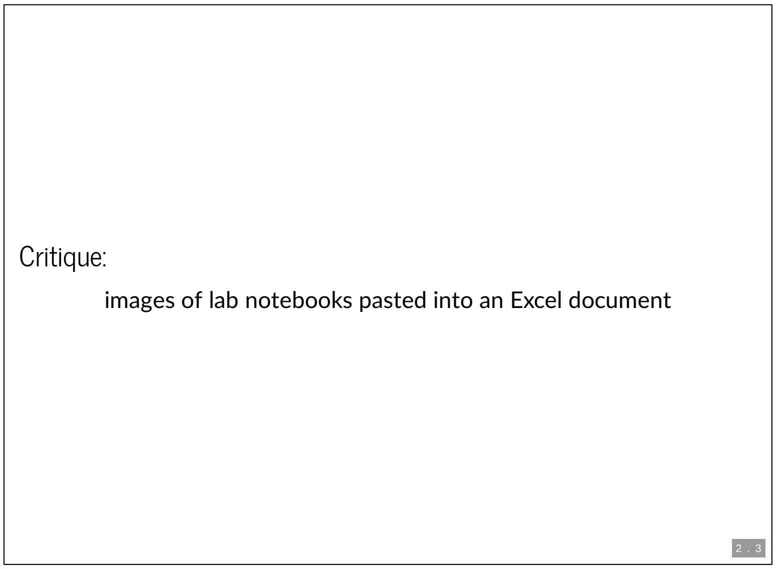
Peter Ralph

13 October – Advanced Biological Statistics

Tidy data

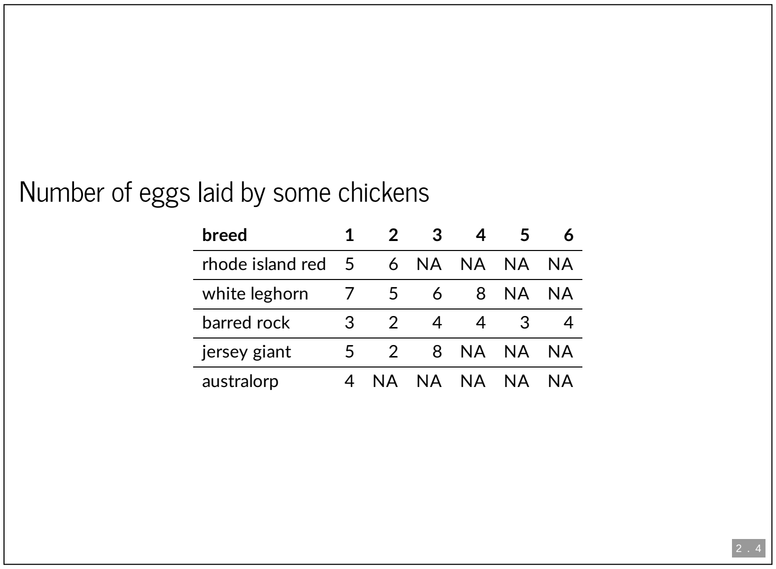
Checklist for data tidiness

- Store a copy of data in a nonproprietary format, (e.g. plain ASCII text)
- Leave an uncorrected file when doing analyses
- Use descriptive names for your data files and variables
- Include a header line with descriptive variable names
- Maintain effective metadata about the data (a README)
- Add new observations to a dataset by *row*
- Add new variables to a dataset by *column*
- A column of data should contain only one data type
- All measurements of the same type should be in the same column



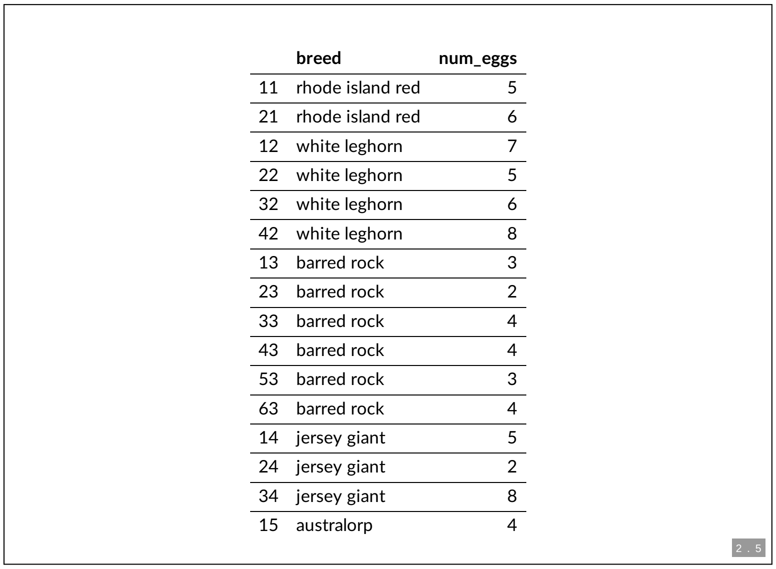
Critique:

images of lab notebooks pasted into an Excel document



Number of eggs laid by some chickens

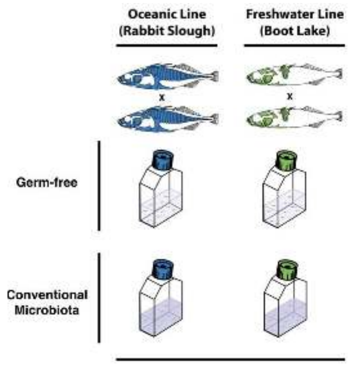
breed	1	2	3	4	5	6
rhode island red	5	6	NA	NA	NA	NA
white leghorn	7	5	6	8	NA	NA
barred rock	3	2	4	4	3	4
jersey giant	5	2	8	NA	NA	NA
australorp	4	NA	NA	NA	NA	NA



	breed	num_eggs
11	rhode island red	5
21	rhode island red	6
12	white leghorn	7
22	white leghorn	5
32	white leghorn	6
42	white leghorn	8
13	barred rock	3
23	barred rock	2
33	barred rock	4
43	barred rock	4
53	barred rock	3
63	barred rock	4
14	jersey giant	5
24	jersey giant	2
34	jersey giant	8
15	australorp	4

Exercise

Design a tidy data format for the stickleback experiment: two strains of stickleback were made microbe free, placed in tanks and either innoculated with microbes or not, then had their gene expression measured with RNA-seq. Sex is recorded, also.



Tools for tidy data

Tidying data is *hard*!

Tools for tidy data

Tidying data is *hard*!

... and often requires expert input.

Tools for tidy data

Tidying data is *hard*!

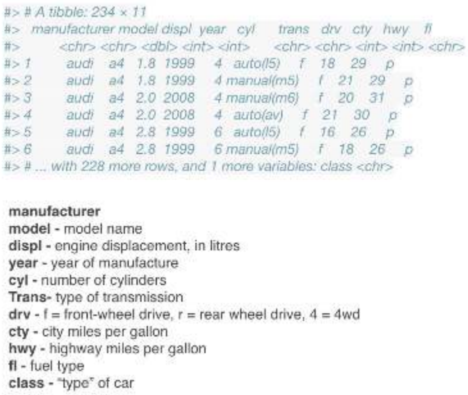
… and often requires expert input.

Many common *data wrangling* operations are made easier by the tidyverse.

The "tidyverse"

- packages that do many of the same things as base functions in R
- designed to do them more "cleanly"
- also includes `ggplot` (for "Grammar of Graphics")

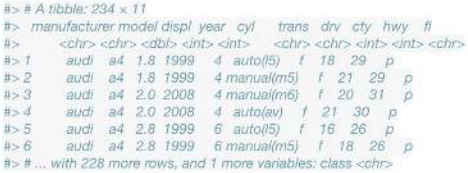
A tibble is a data frame



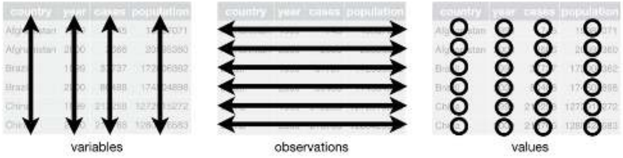
```
#> # A tibble: 234 × 11
#>   manufacturer model displ  year   cyl    trans   drv   cty   hwy    fl
#>          <chr> <chr> <dbl> <int> <int>    <chr> <chr> <int> <int> <chr>
#> 1         audi    a4   1.8  1999     4 auto(l5)     f    18    29     p
#> 2         audi    a4   1.8  1999     4 manual(m5)   f    21    29     p
#> 3         audi    a4   2.0  2008     4 manual(m6)   f    20    31     p
#> 4         audi    a4   2.0  2008     4 auto(av)     f    21    30     p
#> 5         audi    a4   2.8  1999     6 auto(l5)     f    16    26     p
#> 6         audi    a4   2.8  1999     6 manual(m5)   f    18    26     p
#> # ... with 228 more rows, and 1 more variables: class <chr>
```

manufacturer
model - model name
displ - engine displacement, in litres
year - year of manufacture
cyl - number of cylinders
Trans- type of transmission
drv - f = front-wheel drive, r = rear wheel drive, 4 = 4wd
cty - city miles per gallon
hwy - highway miles per gallon
fl - fuel type
class - "type" of car

A tibble is a data frame



```
#> # A tibble: 234 × 11
#>   manufacturer model displ  year   cyl      trans   drv   cty   hwy    fl
#>          <chr> <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>
#> 1         audi    a4   1.8  1999     4   auto(l5)     f    18    29     p
#> 2         audi    a4   1.8  1999     4 manual(m5)     f    21    29     p
#> 3         audi    a4   2.0  2008     4 manual(m6)     f    20    31     p
#> 4         audi    a4   2.0  2008     4   auto(av)     f    21    30     p
#> 5         audi    a4   2.8  1999     6   auto(l5)     f    16    26     p
#> 6         audi    a4   2.8  1999     6 manual(m5)     f    18    26     p
#> # ... with 228 more rows, and 1 more variables: class <chr>
```



Key functions in dplyr

- Pick observations by their values with `filter()`.
- Reorder the rows with `arrange()`.
- Pick variables by their names with `select()`.
- Create new variables with functions of existing variables with `mutate()`.
- Collapse many values down to a single summary with `summarise()`.

`filter()`, `arrange()` and `select()`

```
a1 <- select(airbnb, neighbourhood, price, host_id, beds, bathrooms)
```

```
a2 <- filter(a1, neighbourhood == "Richmond"
                 | neighbourhood == "Woodlawn"
                 | neighbourhood == "Downtown")
```

```
a3 <- arrange(a2, price, neighbourhood)
```

Also, the “pipe”

```
a3 <- (select(airbnb, neighbourhood, price, host_id, beds, bathrooms)
        %>% filter(neighbourhood == "Richmond"
                     | neighbourhood == "Woodlawn"
                     | neighbourhood == "Downtown")
        %>% arrange(price, neighbourhood))
```

Tidyverse:

```
a3 <- (select(airbnb, neighbourhood, price, host_id, beds, bathrooms)
        %>% filter(neighbourhood == "Richmond"
                     | neighbourhood == "Woodlawn"
                     | neighbourhood == "Downtown")
        %>% arrange(price, neighbourhood))
```

Tidyverse:

```
a3 <- (select(airbnb, neighbourhood, price, host_id, beds, bathrooms)
        %>% filter(neighbourhood == "Richmond"
                     | neighbourhood == "Woodlawn"
                     | neighbourhood == "Downtown")
        %>% arrange(price, neighbourhood))
```

Base:

```
a1 <- airbnb[,c("neighbourhood", "price", "host_id", "beds", "bathrooms")]
a2 <- subset(a1, neighbourhood %in% c("Richmond", "Woodlawn", "Downtown"))
a3 <- a2[order(a2$price, a2$price), ]
```

`mutate()` and `transmutate()`

Add new variables:

```
mutate(a3,
    price_per_bed = price / beds,
    price_per_bath = price / bathrooms)
```

Or, make an entirely new data frame:

```
transmute(airbnb,
    price = price,
    price_per_bed = price / beds,
    price_per_bath = price / bathrooms)
```

`group_by()` and `summarize()`

`group_by()` aggregates data by category, e.g.:

```
by_hood <- group_by(a3, neighbourhood)
```

Now, you can calculate *summaries* of other variables *within* each group, e.g.:

```
summarise(by_hood, price = mean(price, na.rm = TRUE))
```

Your turn

1. Make a data frame only including rooms in the top ten neighbourhoods. Then, using only these neighbourhoods…
2. Find the mean `price`, `cleaning_fee`, and ratio of cleaning fee to price, by neighbourhood.
3. Edit your code in (2) to add variables for the 25% and 75% quantile of `price` (use `quantile( )`).
4. Do as in (2) and (3) but splitting by both `neighbourhood` and `room_type` (e.g., finding the mean price of private rooms in Woodlawn).
5. Edit your code in (1) to add a new variable giving the number of characters in the `house_rules` (use `nchar( )`).

Only top ten neighbourhoods

```
neighbourhood_counts <- (airbnb
                         %>% group_by(neighbourhood) %>% summarize(count=n())
                         %>% arrange(count))
top_ten <- neighbourhood_counts$neighbourhood[nrow(neighbourhood_counts) - 0:9]
sub_bnb <- filter(airbnb, neighbourhood %in% top_ten) %>% droplevels
```

Find mean price, cleaning fee, and ratio by hood

```
cleaning <- (sub_bnb
                 %>% group_by(neighbourhood)
                 %>% summarise(mean_price=mean(price, na.rm=TRUE),
                               mean_cleaning_fee=mean(cleaning_fee, na.rm=TRUE),
                               prop_cleaning=mean(cleaning_fee/price, na.rm=TRUE)))
cleaning
```

```
## # A tibble: 10 x 4
##    neighbourhood      mean_price mean_cleaning_fee prop_cleaning
##    <fct>                   <dbl>             <dbl>         <dbl>
##  1 Boise-Eliot              118.              62.6         0.545
##  2 Buckman                  129.              58.6         0.482
##  3 Concordia                113.              55.8         0.519
##  4 Downtown                 237.              85.5         0.466
##  5 Hosford-Abernethy        133.              58.8         0.441
##  6 King                     121.              60.8         0.612
##  7 Northwest District       142.              65.5         0.506
##  8 Overlook                 105.              54.7         0.534
##  9 Richmond                 118.              59.7         0.512
## 10 Sunnyside                114.              56.8         0.513
```

Add quartiles

```
cleaning <- (sub_bnb
                 %>% group_by(neighbourhood)
                 %>% summarise(mean_price=mean(price, na.rm=TRUE),
                               first_quartile_price=quantile(price, probs=0.25,
        na.rm=TRUE),
                               third_quartile_price=quantile(price, probs=0.75,
        na.rm=TRUE),
                               mean_cleaning_fee=mean(cleaning_fee, na.rm=TRUE),
                               prop_cleaning=mean(cleaning_fee/price, na.rm=TRUE)))
cleaning
```

```
## # A tibble: 10 x 6
##    neighbourhood      mean_price first_quartile_price third_quartile_price mean_clea
##    <fct>                   <dbl>                <dbl>                <dbl>
##  1 Boise-Eliot              118.                 75                   135
##  2 Buckman                  129.                 84.5                 146
##  3 Concordia                113.                 70                   130.
##  4 Downtown                 237.                101                   300
##  5 Hosford-Abernethy        133.                 84                   153
##  6 King                     121.                 69                   138
##  7 Northwest District       142.                 89                   163
##  8 Overlook                 105.                 65.8                 119.
##  9 Richmond                 118.                 75                   129
## 10 Sunnyside                114.                 73.2                 134.
```

split also by room type

```
cleaning <- (sub_bnb
                 %>% group_by(neighbourhood, room_type)
                 %>% summarise(mean_price=mean(price, na.rm=TRUE),
                               first_quartile_price=quantile(price, probs=0.25,
        na.rm=TRUE),
                               third_quartile_price=quantile(price, probs=0.75,
        na.rm=TRUE),
                               mean_cleaning_fee=mean(cleaning_fee, na.rm=TRUE),
                               prop_cleaning=mean(cleaning_fee/price, na.rm=TRUE)))
cleaning
```

```
## # A tibble: 25 x 7
## # Groups:   neighbourhood [10]
##    neighbourhood room_type       mean_price first_quartile_price third_quartile_pric
##    <fct>         <fct>                <dbl>                <dbl>                <dbl
##  1 Boise-Eliot   Entire home/apt      130.                    89                 150.
##  2 Boise-Eliot   Private room          74.3                   55                  85
##  3 Boise-Eliot   Shared room           81.8                   36                  75
##  4 Buckman       Entire home/apt      136.                    95                 150.
##  5 Buckman       Private room         106.                    50                  91.
##  6 Buckman       Shared room           25                     25                  25
##  7 Concordia     Entire home/apt      125.                    85                 145
##  8 Concordia     Private room          70.6                   49                  77
##  9 Downtown      Entire home/apt      258.                   125                 300
## 10 Downtown      Private room          95.4                   59                 100
## # … with 15 more rows
```

```
// reveal.js plugins
```

/