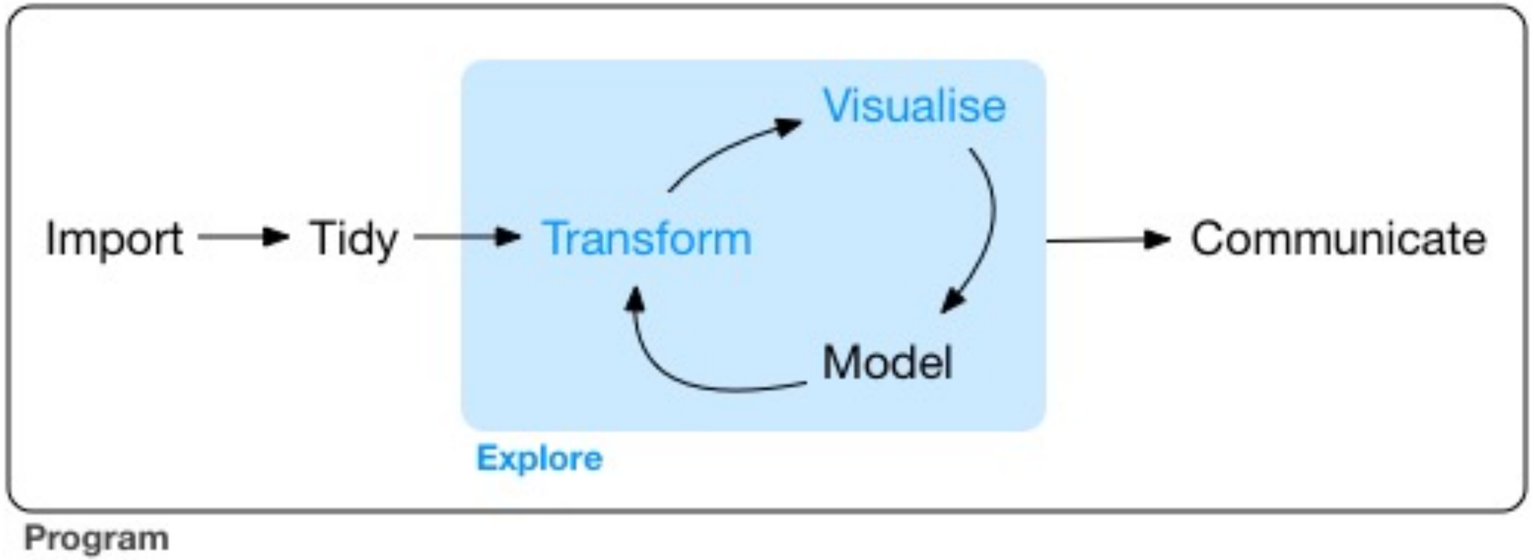
Data Wrangling and Tidyverse Tools

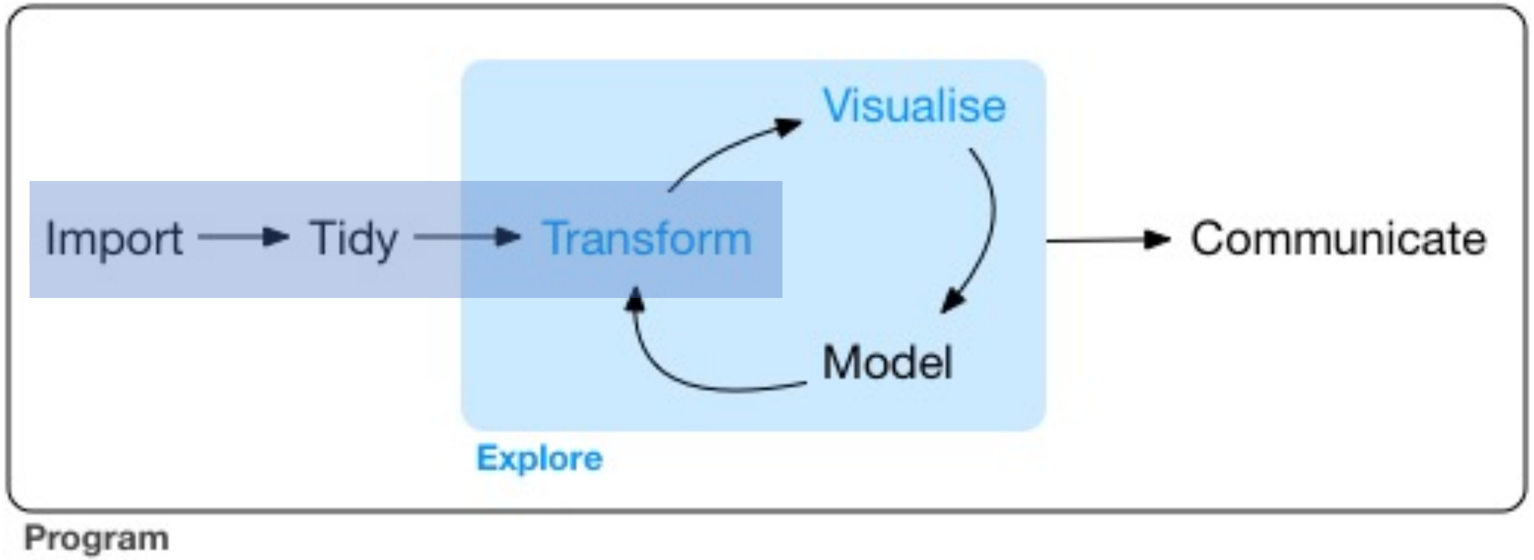
January 26, 2022

Data science workflow



Wickham and Grolemund rhds.had.co.nz

Data science workflow



Wickham and Grolemund rhds.had.co.nz

Don't edit/modify primary data to the extent possible

- Almost anything you need to do can be done in R (or Python or whatever)

- This creates a trail of how you dealt with your data – missing values, labels for cases, etc

- If you have to do this again because you get more or updated data, everything is in place
  - (dealing with errors in raw data?)

Reading in data

- R is not great for data entry
  - Base R commands for importing text data
  - `read.table()`,`read.csv()`
  - menu command in R Studio (avoid?)

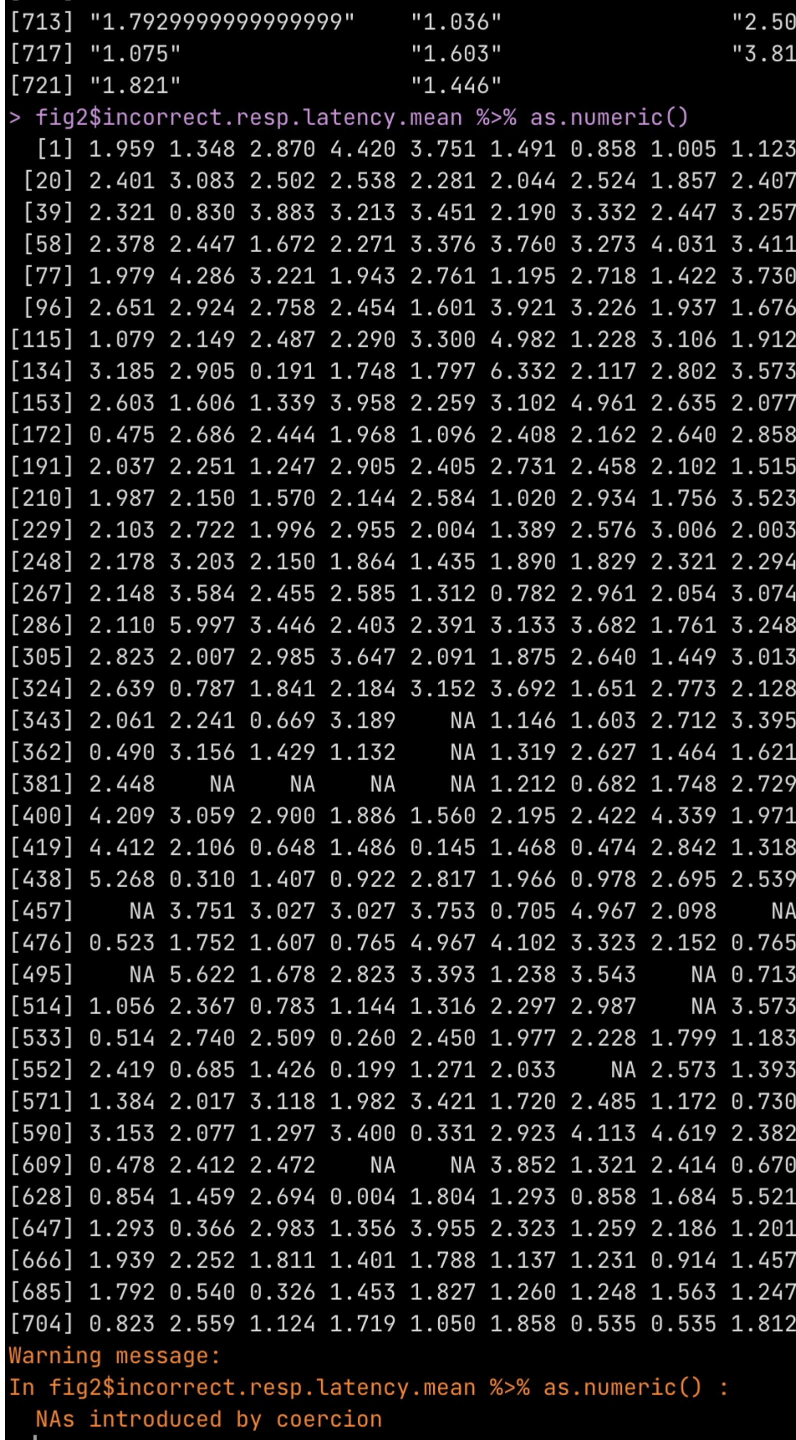
- Read data from Excel spreadsheet (`readxl`)
  - export from Numbers to Excel or CSV

- Read data from other analysis programs like SPSS or SAS (`haven`)

- Read Prism files! (`pzfx`)

What might happen when you read in data

- Check whether read in as character or numeric
- Empty cells in an Excel sheet will cause a numeric variable to be read as character. Easy to fix with as.numeric()
- Old versions of R had a "strings as factors" default setting with undesirable behavior but that is gone now
- Character variables get treated as nominal but you may want to set as factor to change ordering on plots etc



```
[713] "1.7929999999999999"    "1.036"                "2.50
[717] "1.075"                 "1.603"                "3.81
[721] "1.821"                 "1.446"
> fig2$incorrect.resp.latency.mean %>% as.numeric()
  [1] 1.959 1.348 2.870 4.420 3.751 1.491 0.858 1.005 1.123
 [20] 2.401 3.083 2.502 2.538 2.281 2.044 2.524 1.857 2.407
 [39] 2.321 0.830 3.883 3.213 3.451 2.190 3.332 2.447 3.257
 [58] 2.378 2.447 1.672 2.271 3.376 3.760 3.273 4.031 3.411
 [77] 1.979 4.286 3.221 1.943 2.761 1.195 2.718 1.422 3.730
 [96] 2.651 2.924 2.758 2.454 1.601 3.921 3.226 1.937 1.676
[115] 1.079 2.149 2.487 2.290 3.300 4.982 1.228 3.106 1.912
[134] 3.185 2.905 0.191 1.748 1.797 6.332 2.117 2.802 3.573
[153] 2.603 1.606 1.339 3.958 2.259 3.102 4.961 2.635 2.077
[172] 0.475 2.686 2.444 1.968 1.096 2.408 2.162 2.640 2.858
[191] 2.037 2.251 1.247 2.905 2.405 2.731 2.458 2.102 1.515
[210] 1.987 2.150 1.570 2.144 2.584 1.020 2.934 1.756 3.523
[229] 2.103 2.722 1.996 2.955 2.004 1.389 2.576 3.006 2.003
[248] 2.178 3.203 2.150 1.864 1.435 1.890 1.829 2.321 2.294
[267] 2.148 3.584 2.455 2.585 1.312 0.782 2.961 2.054 3.074
[286] 2.110 5.997 3.446 2.403 2.391 3.133 3.682 1.761 3.248
[305] 2.823 2.007 2.985 3.647 2.091 1.875 2.640 1.449 3.013
[324] 2.639 0.787 1.841 2.184 3.152 3.692 1.651 2.773 2.128
[343] 2.061 2.241 0.669 3.189    NA 1.146 1.603 2.712 3.395
[362] 0.490 3.156 1.429 1.132    NA 1.319 2.627 1.464 1.621
[381] 2.448    NA    NA    NA    NA 1.212 0.682 1.748 2.729
[400] 4.209 3.059 2.900 1.886 1.560 2.195 2.422 4.339 1.971
[419] 4.412 2.106 0.648 1.486 0.145 1.468 0.474 2.842 1.318
[438] 5.268 0.310 1.407 0.922 2.817 1.966 0.978 2.695 2.539
[457]    NA 3.751 3.027 3.027 3.753 0.705 4.967 2.098    NA
[476] 0.523 1.752 1.607 0.765 4.967 4.102 3.323 2.152 0.765
[495]    NA 5.622 1.678 2.823 3.393 1.238 3.543    NA 0.713
[514] 1.056 2.367 0.783 1.144 1.316 2.297 2.987    NA 3.573
[533] 0.514 2.740 2.509 0.260 2.450 1.977 2.228 1.799 1.183
[552] 2.419 0.685 1.426 0.199 1.271 2.033    NA 2.573 1.393
[571] 1.384 2.017 3.118 1.982 3.421 1.720 2.485 1.172 0.730
[590] 3.153 2.077 1.297 3.400 0.331 2.923 4.113 4.619 2.382
[609] 0.478 2.412 2.472    NA    NA 3.852 1.321 2.414 0.670
[628] 0.854 1.459 2.694 0.004 1.804 1.293 0.858 1.684 5.521
[647] 1.293 0.366 2.983 1.356 3.955 2.323 1.259 2.186 1.201
[666] 1.939 2.252 1.811 1.401 1.788 1.137 1.231 0.914 1.457
[685] 1.792 0.540 0.326 1.453 1.827 1.260 1.248 1.563 1.247
[704] 0.823 2.559 1.124 1.719 1.050 1.858 0.535 0.535 1.812
Warning message:
In fig2$incorrect.resp.latency.mean %>% as.numeric() :
  NAs introduced by coercion
```

Tidyverse vs base R

- Suite of packages to make R code more readable
  - always more than one way to do things in R!!

- Minimize use of $ (`data_set$variable_1`)

- Introduced "pipe" operator
  - %>%
  - thing on left side of %>% becomes first argument of thing on right side
  - `function(x, y)` and `x %>% function(y)` are equivalent
  - base R now incorporates similar operator as of version 4.1 |>

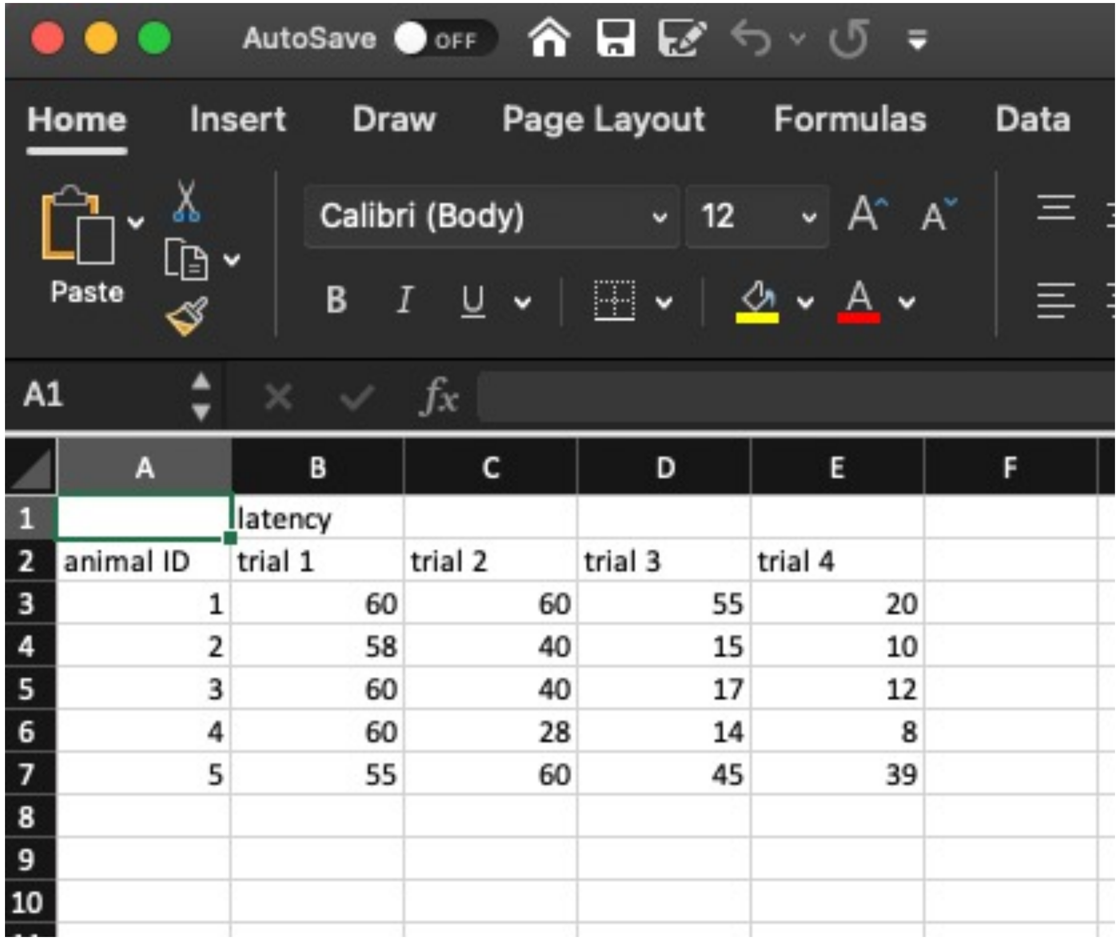
```
read_excel()

glimpse()

mutate(), summarize(), filter(), select()

pivot_longer(), pivot_wider()

ggplot()
    ggplot(aes(x = x_variable, y = y_variable) +
    geom_point() + ...
```



	A	B	C	D	E	F
1		latency				
2	animal ID	trial 1	trial 2	trial 3	trial 4	
3	1	60	60	55	20	
4	2	58	40	15	10	
5	3	60	40	17	12	
6	4	60	28	14	8	
7	5	55	60	45	39	
8						
9						
10						

20220126_demo.R