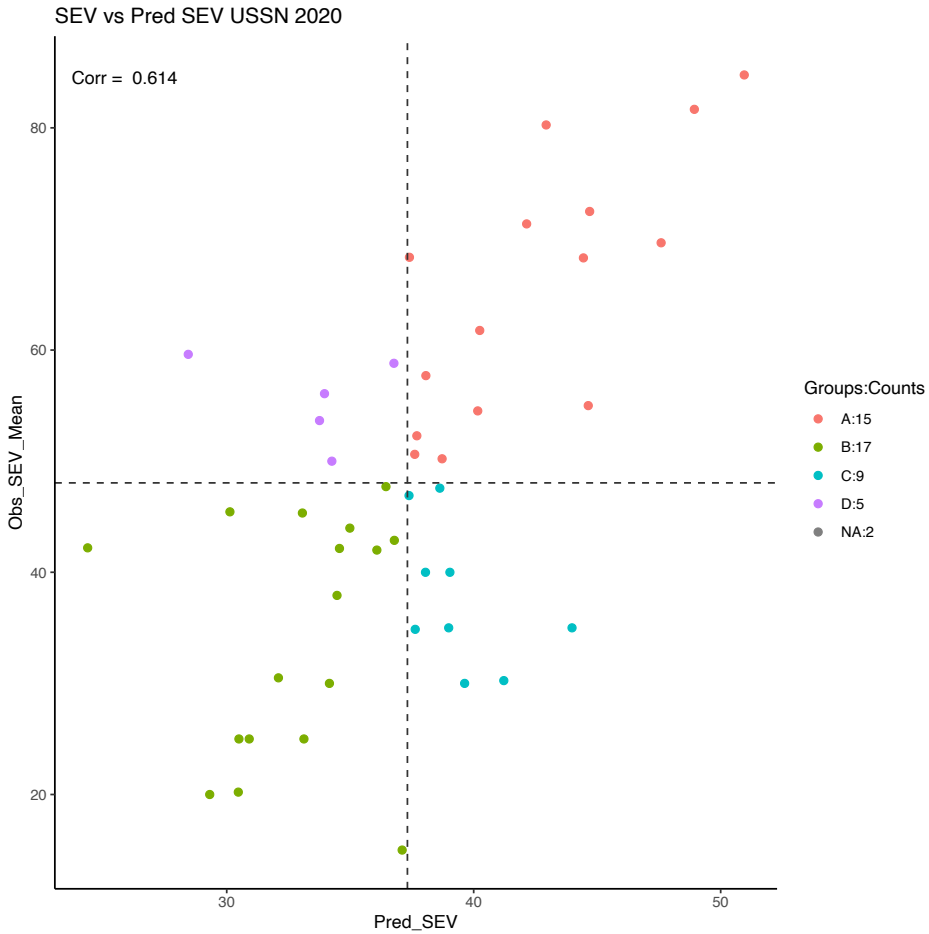
Data cleaning with R

Jeanette Lyerly

Wheat CAP Genomic Selection Workshop

July 2023, Raleigh, NC



NC STATE UNIVERSITY

Background

- About SunGrains
  - Cooperative program with seven southern universities and the Eastern Regional Genotyping Center
- Implement genomic selection for southern wheat breeding
  - Pool resources and data
  - Prediction modeling for new breeding material
- Incoming data from a variety of sources

Why do we need a data plan?

Use the information you have gathered

Share information with collaborators

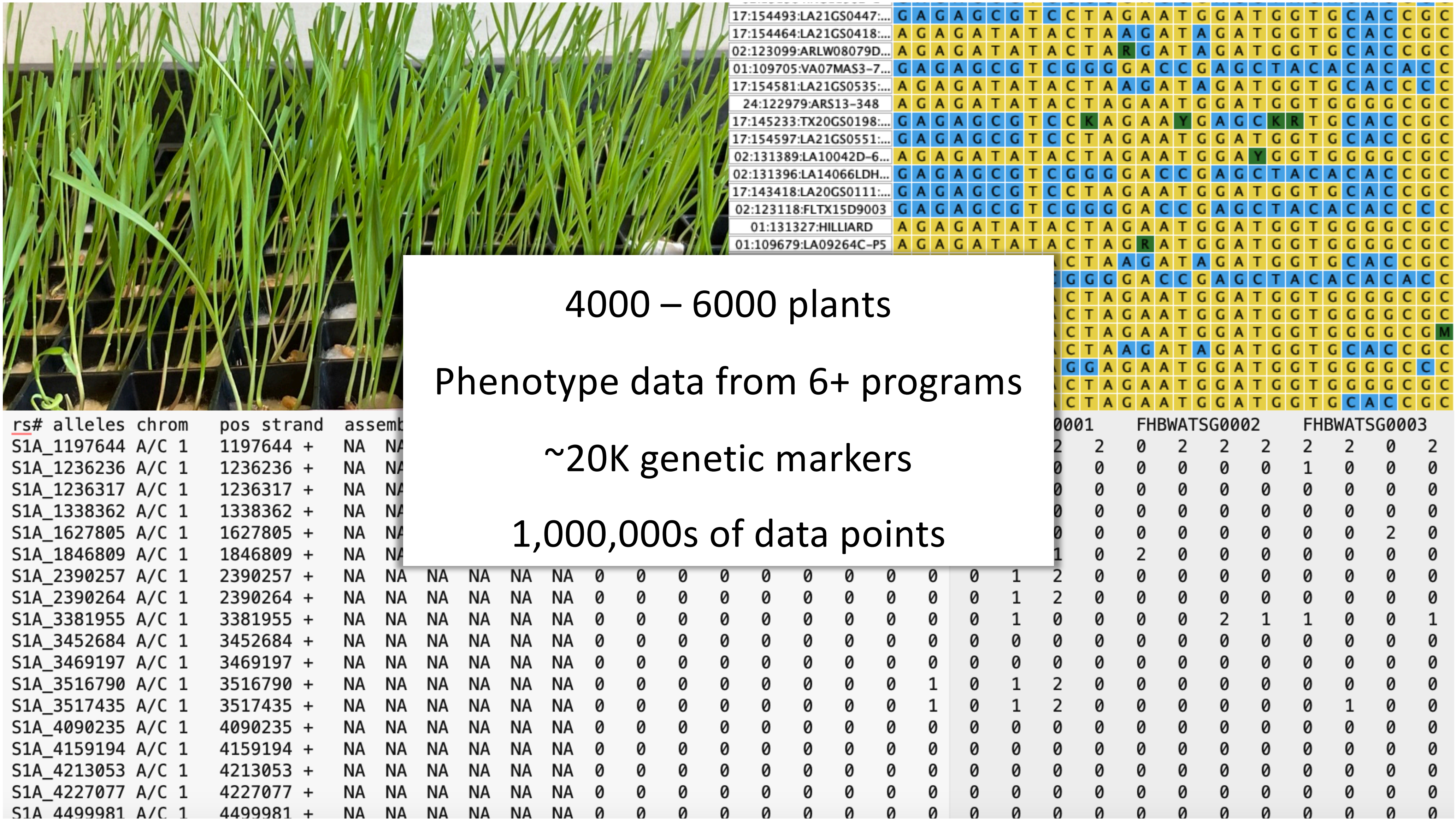
Reproducibility

Avoid making more work for yourself later

Types of data

- Genotype/GBS data
- Phenotype data
- Marker data
- Data from databases or other resources (climate data, etc.)

SNP	S1A_1158055	S1A_1197644	FHB Severity	Height
Chromosome	1	1		
Position	1158055	1197644		
Genotype				
BESS	0	2	22.24	36.16
COKER9835	0	2	59.74	31.23
JAMESTOWN	0	0	27.04	32.44
NC05-15-99	0	0	23.68	45.52
NC05-39-314	0	0	43.40	34.76
NC08-23324	0	1	39.16	34.22
NC09-20768	0	0	47.53	32.29
NC09-20986	0	0	18.98	33.26
NC09-22422	0	1	24.31	32.15
NC10435-11	0	2	32.96	32.30
NC10-23663	0	0	.	.
NC10-23720	1	1	.	.
NC10014-52	0	0	.	.
NC11-21982	0	2	.	.
NC11-22289	0	0	.	.
NC11-23321	0	0	.	.
NC11359-25	0	1	.	.
NC11360-42	0	1	.	.
NC11362-343	0	0	.	.
NC11363-86	0	0	.	.
NC12-20661	0	2	.	.
NC8170-4-3	0	2	.	.


17:154493:LA21GS0447:... G A G A G C G T C C T A G A A T G G A T G G T G C A C C G C
17:154464:LA21GS0418:... A G A G A T A T A C T A A G A T A G A T G G T G C A C C G C
02:123099:ARLW08079D... A G A G A T A T A C T A R G A T A G A T G G T G C A C C G C
01:109705:VA07MAS3-7... G A G A G C G T C G G G G A C C G A G C T A C A C A C A C C
17:154581:LA21GS0535:... A G A G A T A T A C T A A G A T A G A T G G T G C A C C C C
24:122979:ARS13-348 A G A G A T A T A C T A G A A T G G A T G G T G G G G C G C
17:145233:TX20GS0198:... G A G A G C G T C C K A G A A Y G A G C K R T G C A C C G C
17:154597:LA21GS0551:... G A G A G C G T C C T A G A A T G G A T G G T G C A C C G C
02:131389:LA10042D-6... A G A G A T A T A C T A G A A T G G A Y G G T G G G G C G C
02:131396:LA14066LDH... G A G A G C G T C G G G G A C C G A G C T A C A C A C C G C
17:143418:LA20GS0111:... G A G A G C G T C C T A G A A T G G A T G G T G C A C C G C
02:123118:FLTX15D9003 G A G A G C G T C G G G G A C C G A G C T A C A C A C C C C
01:131327:HILLIARD A G A G A T A T A C T A G A A T G G A T G G T G G G G C G C
01:109679:LA09264C-P5 A G A G A T A T A C T A G R A T G G A T G G T G G G G C G C
rs# alleles chrom pos strand assemb 0001 FHBWATSG0002 FHBWATSG0003
S1A_1197644 A/C 1 1197644 + NA NA 2 2 0 2 2 2 2 2 0 2
S1A_1236236 A/C 1 1236236 + NA NA 0 0 0 0 0 0 1 0 0 0
S1A_1236317 A/C 1 1236317 + NA NA 0 0 0 0 0 0 0 0 0 0
S1A_1338362 A/C 1 1338362 + NA NA 0 0 0 0 0 0 0 0 0 0
S1A_1627805 A/C 1 1627805 + NA NA 0 0 0 0 0 0 0 0 2 0
S1A_1846809 A/C 1 1846809 + NA NA 1 0 2 0 0 0 0 0 0 0
S1A_2390257 A/C 1 2390257 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 1 2 0 0 0 0 0 0 0 0 0
S1A_2390264 A/C 1 2390264 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 1 2 0 0 0 0 0 0 0 0 0
S1A_3381955 A/C 1 3381955 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 2 1 1 0 0 1
S1A_3452684 A/C 1 3452684 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
S1A_3469197 A/C 1 3469197 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
S1A_3516790 A/C 1 3516790 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 1 0 1 2 0 0 0 0 0 0 0 0 0
S1A_3517435 A/C 1 3517435 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 1 0 1 2 0 0 0 0 0 0 1 0 0
S1A_4090235 A/C 1 4090235 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
S1A_4159194 A/C 1 4159194 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
S1A_4213053 A/C 1 4213053 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
S1A_4227077 A/C 1 4227077 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
S1A_4499981 A/C 1 4499981 + NA NA NA NA NA NA 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

4000 – 6000 plants

Phenotype data from 6+ programs

~20K genetic markers

1,000,000s of data points

Consider your data plan early

- Thinking through your data at the beginning of your project can save you time later

Things to consider in your data plan

- How will data be collected?
  - Account for multiple people/programs?
- What do you need to do with this data?
- How will data be stored?
- Who will have access to the data?
  - Account for privacy, proprietary data
- How will data and results be distributed?
  - Format in = format out?

Common issues

- Naming conventions
  - Differences in line names, abbreviations
    - NC-1234 vs NC 1234
  - Different names for the same trait
    - YLD vs Yield
- Units and scales
- Duplicate records
- Other glitches
  - Excel reformats data as a date, integers become strings, etc.

Data collection and formatting

- **Pick a way!**
- Consistency across experiments
  - Makes data easier to process
  - Makes data easier to format for other applications

EXPT	YR	LOC	ENTRY	ID	YLD_BUPA	TW_LBBU	HD_YR
GWN	23	NCK	001	AGS8417	68.9	50.9	85
GWN	23	NCK	002	GA149D7-8-35	63.0	57.6	90
GWN	23	NCK	003	GA178X4-ID-8-15	55.9	61.2	88
GWN	23	NCK	004	X14-047-178-14-7	46.0	49.4	91
GWN	23	NCK	005	LA1215715CB	49.0	59.2	88
GWN	23	NCK	006	NC1474A-157	60.0	56.4	90
GWN	23	NCK	007	NC146LDH-44	75.9	57.5	91
GWN	23	NCK	008	NC17847-48	73.1	58.1	94
GWN	23	NCK	009	NC154-15971	60.0	59.6	90
GWN	23	NCK	010	TX054D70A	48.7	57.4	85
GWN	23	NCK	011	AGS7218	67.0	52.7	82
GWN	23	NCK	012	GA179D7-8-42	61.1	59.5	87
GWN	23	NCK	013	GA18M5-ID-8-13	54.0	63.0	85
GWN	23	NCK	014	X12-048-179-14-8	44.1	51.3	88
GWN	23	NCK	015	LA1315718CL	47.1	61.0	85
GWN	23	NCK	016	NC1575A-152	58.1	58.2	87
GWN	23	NCK	017	NC116LDH-67	74.0	59.3	88
GWN	23	NCK	018	NC18472-77	71.2	59.9	91
GWN	23	NCK	019	NC164-15432	58.0	61.5	87
GWN	23	NCK	020	TX074A80X	46.8	59.3	82

Name	entry	DYLD	TW
Hilliard	1	44	61.2
AGS3030	2	47	62.8

	A	B	C	D	E	F	G
1	plot	bloc	entry	name	heading	tw	yield
2	1	1	1	Hilliard	99	58.3	66.4
3	2	1	2	AGS3030	97	58.4	62.1

	GAWN-PL19			Mean	Mean
				82.8	57.1
Entry	Line	Pedigree	% of Mean	BUPA avg.	Test Wt. avg.
1	Hilliard		99.4	82.3 *	46.8
2	AGS3030		101.5	84	58.4

ENT	LOC	GENO	PED	R1_YIELD	R2_YIELD	R1_TWT	R2_TWT	% R1_MST	% R2_MST
1	FL	Hilliard		55.3	56.9	42.8	44.3	9.3	9.6
2	FL	AGS3030		71.3	46.8	53.2	33.0	9.9	9.0

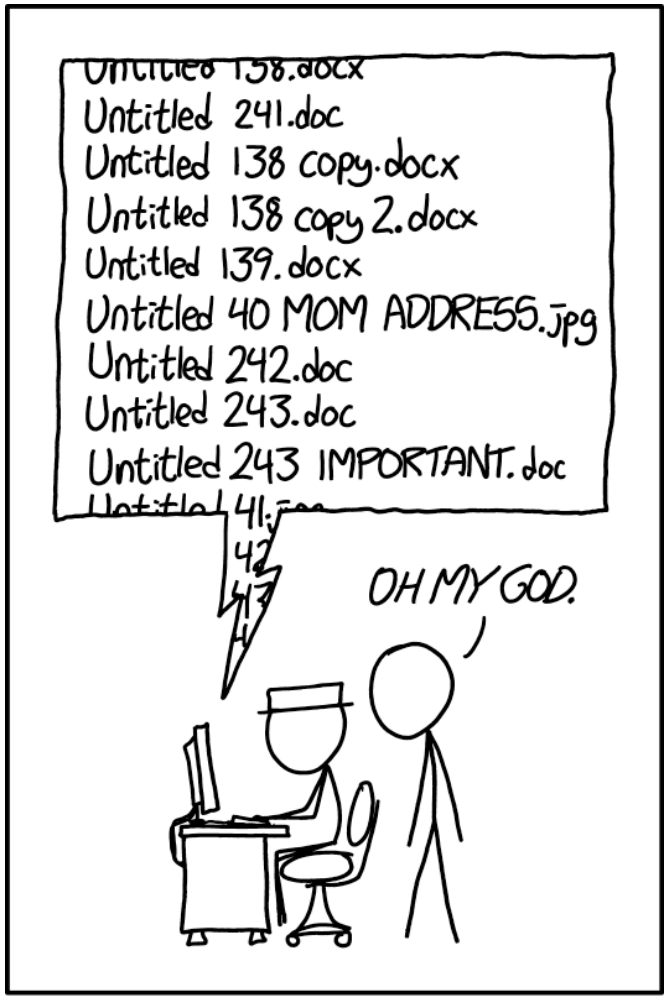
	A	B	C	D	E	F	G	H
1	EXPT	PLOT	REP	ENT	LOC	VAR	BUPA	TWT
2	GAWN19WN	247	2	1	WN	HILLIARD	63.541519	56
3	GAWN19WN	231	2	2	WN	AGS3030	44.979577	53

Data collection and formatting

- Put only one thing in a cell
- Choose good names
  - Avoid spaces and special characters
- Be consistent!
  - Variable names
  - Units
  - Code for missing data (hint – zero is not missing data)

Entry	Variety	Location	PlantHt
1	AB 22866	TOWN, NC	28.0
2	AB21090	TOWN, NC	31.0
3	AB- 23945	TOWN, NC	0
4	AB_21642	TOWN, NC	31.5
5	AB-24757*	TOWN, NC	30.7

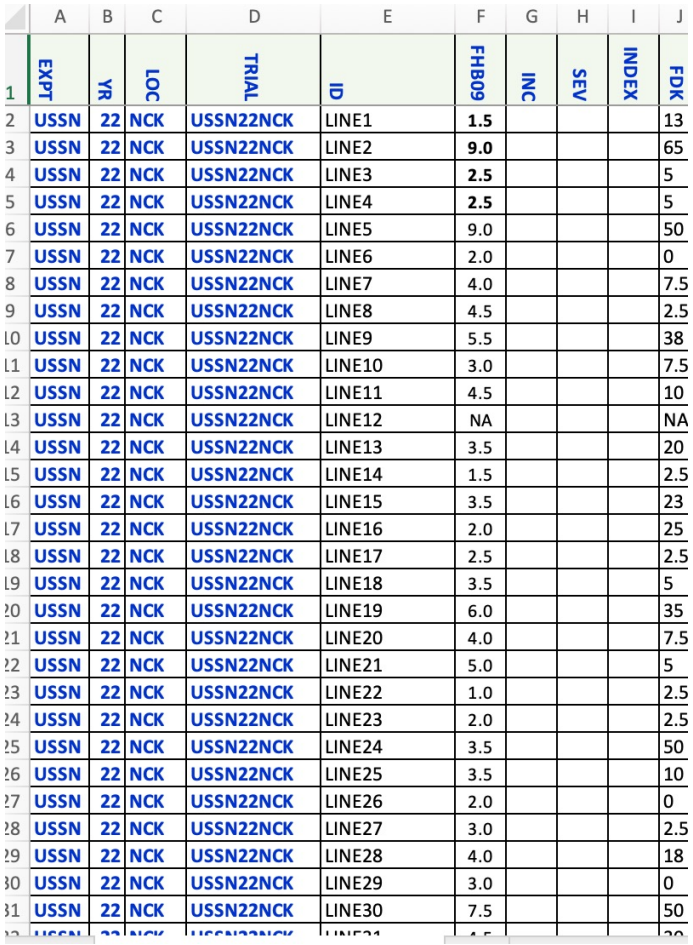
Entry	Variety	Location	State	HT
1	AB-22866	TOWN	NC	71.1
2	AB-21090	TOWN	NC	78.7
3	AB-23945	TOWN	NC	NA
4	AB-21642	TOWN	NC	79.9
5	AB-24757	TOWN	NC	78.1



https://xkcd.com/1459/

File and folder structures

- Think about file names and folder structures
  - Consider your workflow
- Use names that make sense (NOT that only make sense at the time)
  - analysis_final
  - analysis_final2
  - analysis_usethisone
- Use ReadMe or other descriptive files



	A	B	C	D	E	F	G	H	I	J
1	EXPT	YR	LOC	TRIAL	ID	FHB09	INC	SEV	INDEX	FDK
2	USSN	22	NCK	USSN22NCK	LINE1	**1.5**				13
3	USSN	22	NCK	USSN22NCK	LINE2	**9.0**				65
4	USSN	22	NCK	USSN22NCK	LINE3	**2.5**				5
5	USSN	22	NCK	USSN22NCK	LINE4	**2.5**				5
6	USSN	22	NCK	USSN22NCK	LINE5	9.0				50
7	USSN	22	NCK	USSN22NCK	LINE6	2.0				0
8	USSN	22	NCK	USSN22NCK	LINE7	4.0				7.5
9	USSN	22	NCK	USSN22NCK	LINE8	4.5				2.5
10	USSN	22	NCK	USSN22NCK	LINE9	5.5				38
11	USSN	22	NCK	USSN22NCK	LINE10	3.0				7.5
12	USSN	22	NCK	USSN22NCK	LINE11	4.5				10
13	USSN	22	NCK	USSN22NCK	LINE12	NA				NA
14	USSN	22	NCK	USSN22NCK	LINE13	3.5				20
15	USSN	22	NCK	USSN22NCK	LINE14	1.5				2.5
16	USSN	22	NCK	USSN22NCK	LINE15	3.5				23
17	USSN	22	NCK	USSN22NCK	LINE16	2.0				25
18	USSN	22	NCK	USSN22NCK	LINE17	2.5				2.5
19	USSN	22	NCK	USSN22NCK	LINE18	3.5				5
20	USSN	22	NCK	USSN22NCK	LINE19	6.0				35
21	USSN	22	NCK	USSN22NCK	LINE20	4.0				7.5
22	USSN	22	NCK	USSN22NCK	LINE21	5.0				5
23	USSN	22	NCK	USSN22NCK	LINE22	1.0				2.5
24	USSN	22	NCK	USSN22NCK	LINE23	2.0				2.5
25	USSN	22	NCK	USSN22NCK	LINE24	3.5				50
26	USSN	22	NCK	USSN22NCK	LINE25	3.5				10
27	USSN	22	NCK	USSN22NCK	LINE26	2.0				0
28	USSN	22	NCK	USSN22NCK	LINE27	3.0				2.5
29	USSN	22	NCK	USSN22NCK	LINE28	4.0				18
30	USSN	22	NCK	USSN22NCK	LINE29	3.0				0
31	USSN	22	NCK	USSN22NCK	LINE30	7.5				50

Data templates

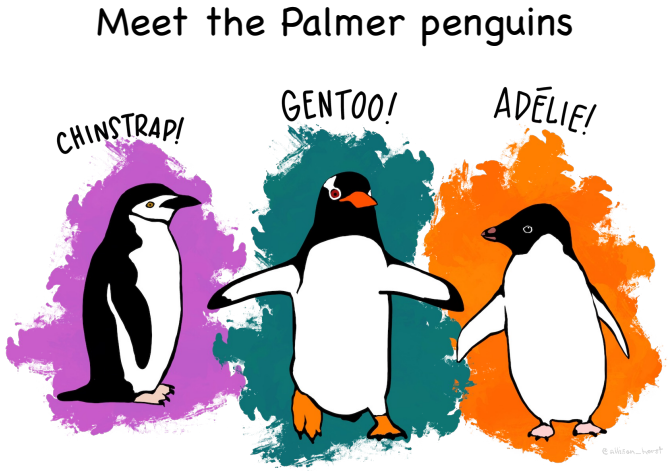
- Using a template can alleviate a lot of issues
  - If you are formatting for submission to a database, then some type of template will be required
- Using a template similar to existing data recording methods
  - May make it easier to use
  - Increase adoption

GS generates a lot of data!

- Field and genetic data from multiple research groups
- Mountains of data!
- Quicky move beyond what you would want to handle in Excel

Beginning to wrangle data in R

- R can have a steep learning curve
  - Know where to find help
  - Incremental improvements are ok
- Practice helps
  - TidyTuesday, Advent of Code, R4DS book clubs, online tutorials
  - Useful to learn skills on data that isn't yours



Artwork by @allison_horst

Horst AM, Hill AP, Gorman KB (2020).

palmerpenguins: Palmer Archipelago (Antarctica) penguin data.

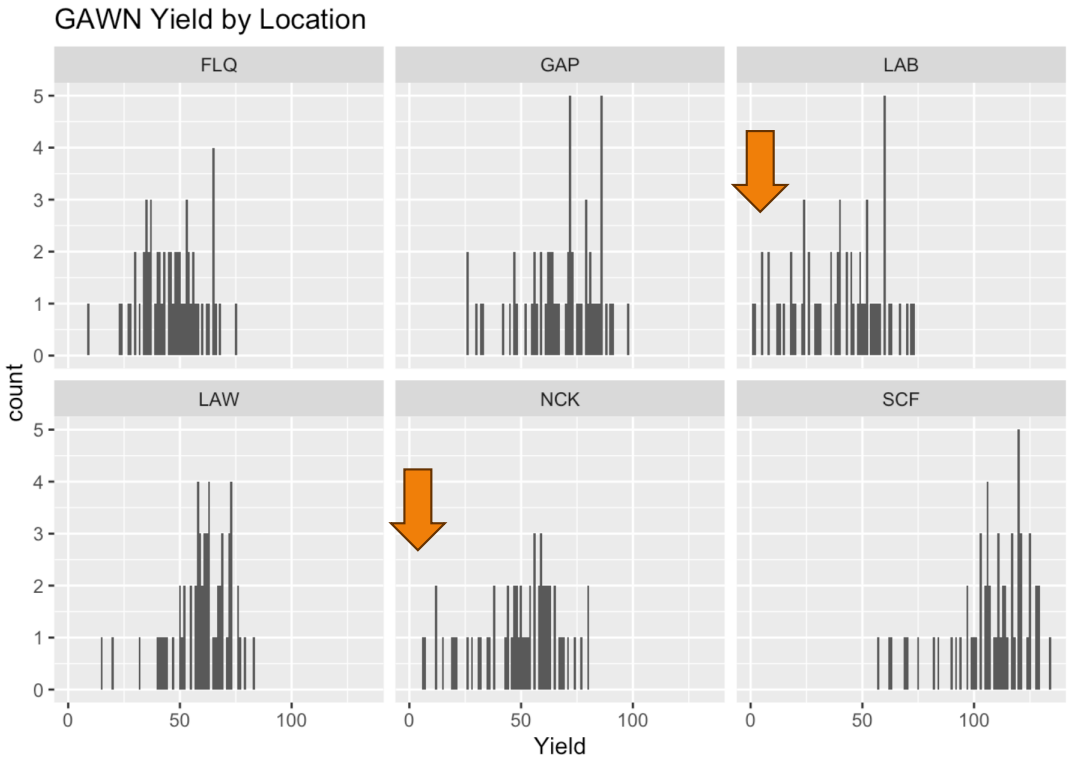
https://allisonhorst.github.io/palmerpenguins/

Wrangling data in R

- No single approach to data cleaning
  - Depends on data type, project, etc.
  - What works best for one may not be best for another
- Packages that can help with data cleaning
  - Tidyverse, janitor
  - R is ever evolving – lots of packages available

Avoiding pitfalls

- Look at your data!
- Exploratory analysis is important
  - See trends
  - Identify issues



Avoiding pitfalls

- Avoid the trap “I will remember this”
- Hint: you will not remember this
  - Make notes about how something is done/calculated
  - It may be some time before you come back to things
- Use the ReadMe files

Avoiding pitfalls

- Comment your code!
  - Comment for yourself - see trap “I will remember this”
  - Comment for your collaborators and/or people who will take over your project

```
68  ggplot(mydata, aes(x = YLD_BUPA)) +
69    geom_histogram(binwidth = 1) + #make a histogram of the yield data
70    facet_wrap(vars(LOC)) + #wrap over locations
71    labs(title = paste(mynursery, "Yield by Location", sep = " "), x = "Yield") #label the plot
```

Hands-on example

- Read in some data
- Check the data
- Create some summaries and do basic visuals
- Do a bit of data cleaning
- Export a clean file for downstream analysis