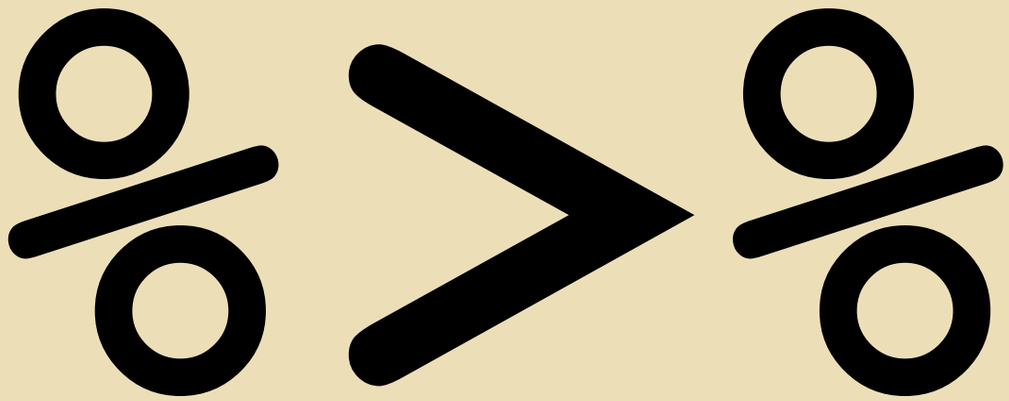
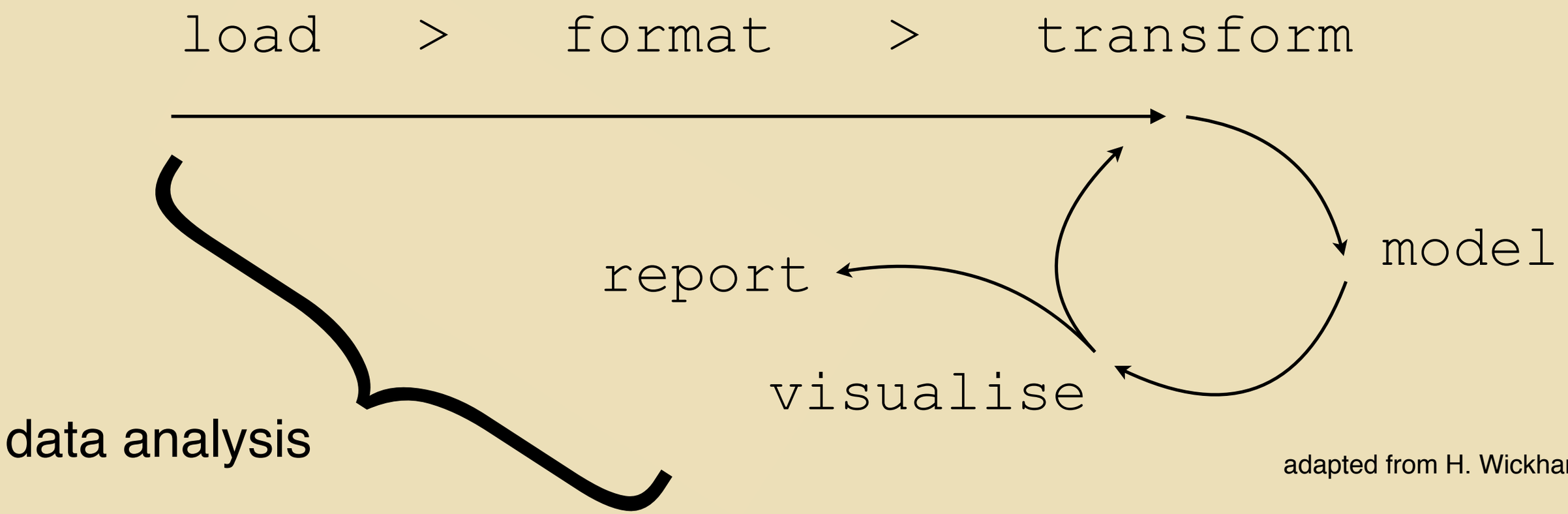


Ceci n'est pas une pipe.

Next generation programming in R

Florian Uhlitz
uhlitz@hu-berlin.de
uhlitz.github.io

data wrangling



adapted from H. Wickham

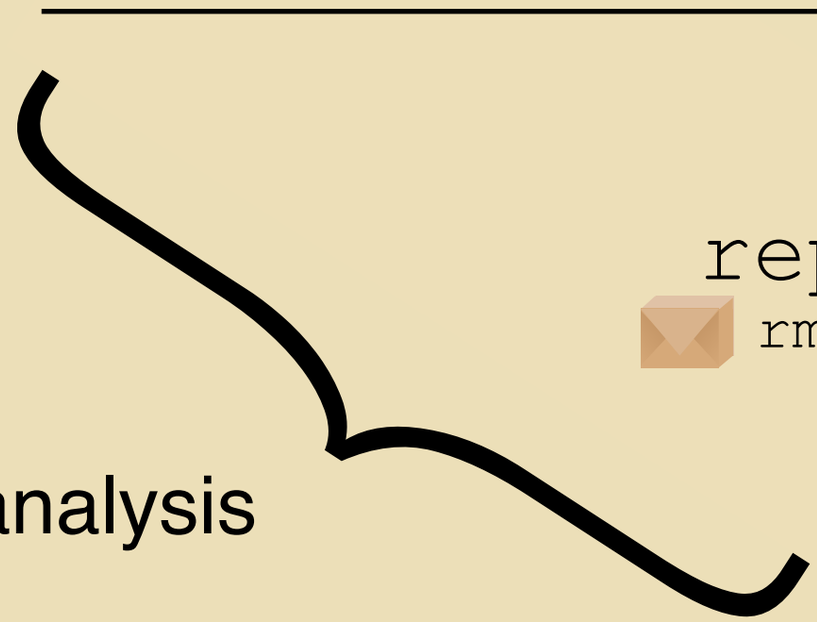
data wrangling

readr

tidyr

dplyr

load %>% format %>% transform



model

base

broom

report

rmarkdown

visualise

ggplot2

data analysis

adapted from H. Wickham

magrittr %>%

magrittr

In a pipe, the result of the left hand statement is handed over to the function on the right hand side:

```
    f(x, y)
⇔ x %>% f(y)
```

```
    f(x, y, z)
⇔ x %>% f(y, z)

    f2(f1(x), y)
⇔ f1(x) %>% f2(y)
```

stolen from

…idea ~~similar to~~ Unix pipe operator |

magrittr

```
summarise(
  merge(
    filter(
      transform(
        raw_interesting_data, somehow
      ),
      the_good_parts
    ),
    lookup_table
  ),
  result = mean(values)
)
```

nested functions

```
raw_interesting_data %>%
  transform(somehow) %>%
  filter(the_good_parts) %>%
  merge(lookup_table) %>%
  summarise(result = mean(values))
```

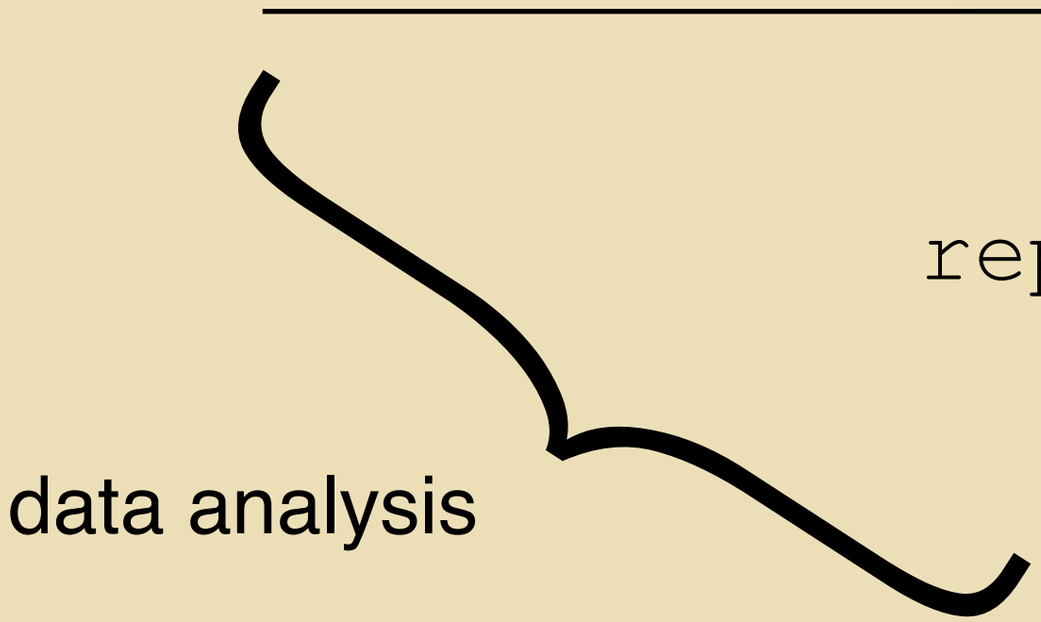

chain of functions

data wrangling

readr tidyr dplyr



load %>% format %>% transform



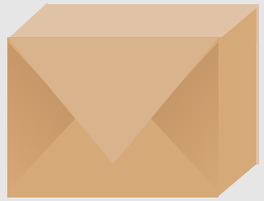
model

report

visualise

data analysis

adapted from H. Wickham



readr, readxl, haven

```
readr::read_csv()
readr::read_tsv()
readr::read_log()
readr::read_delim()
readr::read_fwf()
readr::read_table()

readxl::read_excel()

haven::read_sas()
haven::read_spss()
haven::read_stata()
```

~10x faster than base R functions

+

progress bar

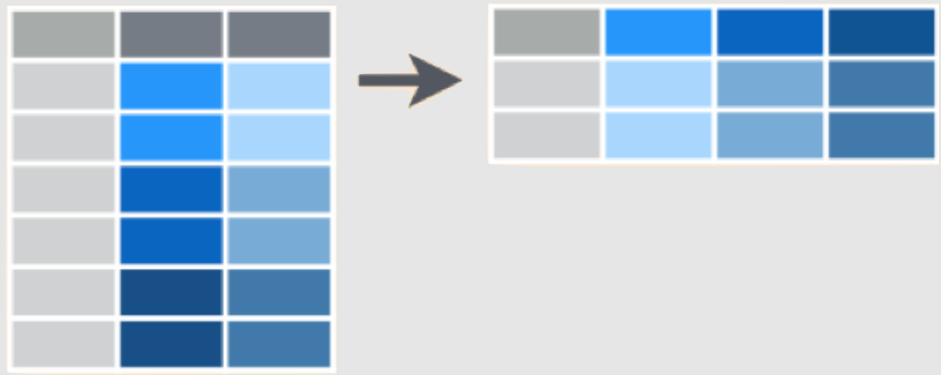
```
> read_tsv("gff3v19.tsv")
|=================================================================|  68%  730 MB
```

tidyr

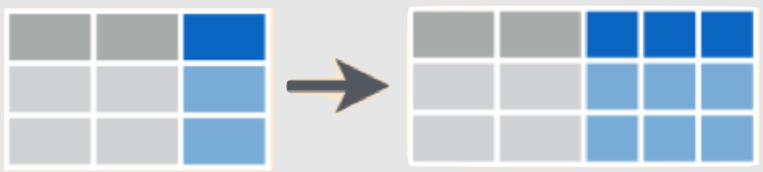
Formatting

`pivot_longer()`

`pivot_wider()`

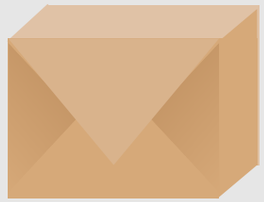


`separate()`



`unite()`

adapted from rstudio.com/resources/cheatsheets/

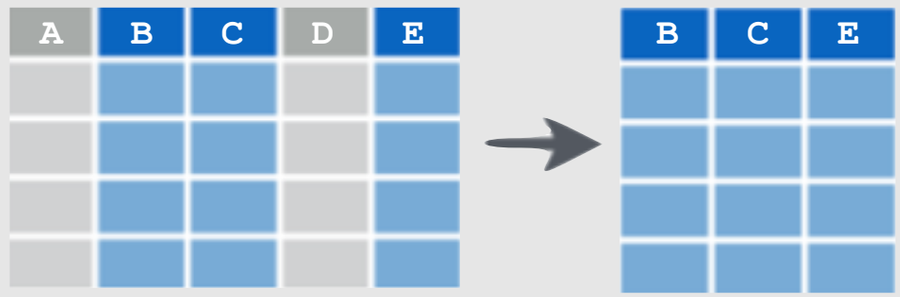


dplyr

Subsetting

`filter(x > 1)`

`select(B, C, E)`



adapted from rstudio.com/resources/cheatsheets/

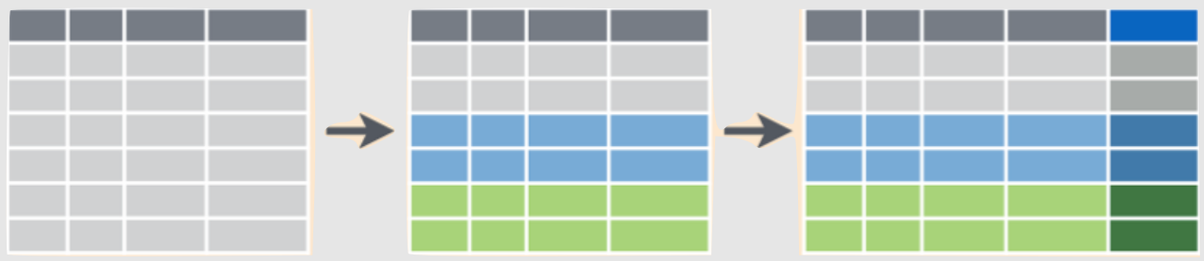
dplyr

Transforming

`mutate(z = x + y)`

x	y			z
1	4			5
2	5			7
3	6			9

`group_by() %>% mutate()`



Summarising

`summarise(A = sum(x), B = sum(y))`

x	y			A	B
1	4			6	15
2	5				
3	6				

`group_by() %>% summarise()`

adapted from rstudio.com/resources/cheatsheets/

What`s tidy data?

KEEP CALM AND TIDY UP

Tidy data principle

»Tidy data sets are all alike; every messy data set is messy in its own way.«

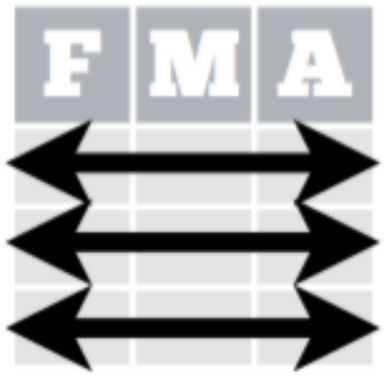
Hadley Wickham

Tidy data definition

In a tidy data set:

Each **variable** is saved in its own **column**

&



Each **observation** is saved in its own **row**

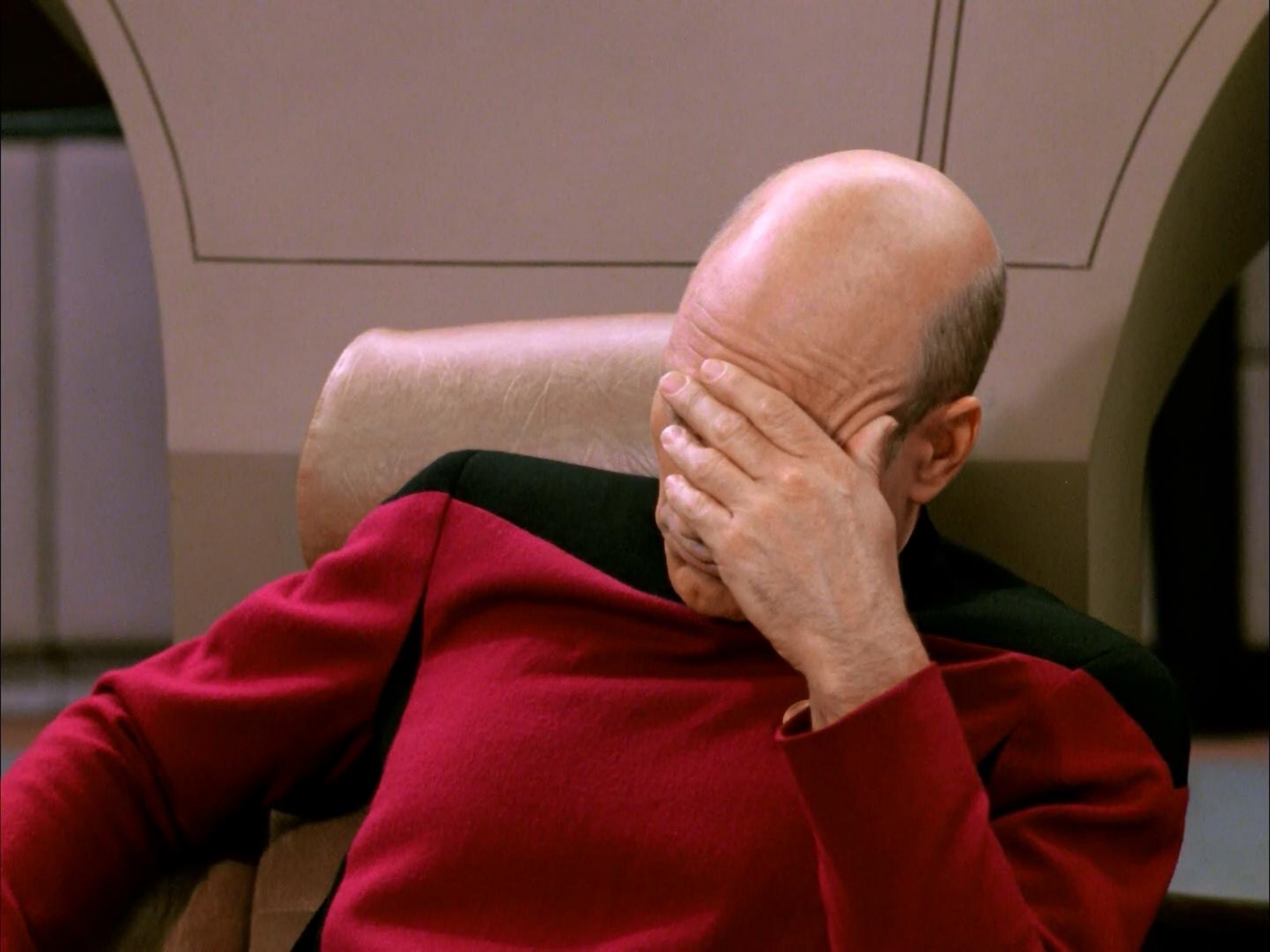
Wickham, H. (2014). Tidy Data. Journal of Statistical Software

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1					max/min	63		7		16		27		12		5		8		4		2		3		3			
2																													
3	Type	Well	minutues			p-Akt		p-c-Jun		p-ERK2		p-GSK3a/B		p-JNK		p-MEK1		p-p70 S6 Kinase		p-Stat3 (Ser727)		p-NF-kB p65		p-p38 MAPK		p-HSP27			
4	X1	A9	60	EGF		10943		17901		127		1221		259		82		1750		98		88		28		63			
5	X2	B9	30	EGF		10283		18950		132		1282		274		70		1795		94		94		28		80			
6	X3	C9	20	EGF		9867		17860		116		1067		229		57		1606		96		88		27		69			
7	X4	D9	10	EGF		6718		19262		146		1226		404		72		1813		106		83		14		80			
8	X5	E9	60	FGF2		2970		17205		106		960		292		58		1595		90		87		22		84			
9	X6	F9	30	FGF2		2462		15594		108		923		215		60		1400		85		65		26		82			
10	X7	G9	20	FGF2		2651		18028		96		844		231		54		1378		85		61		22		66			
11	X8	H9	10	FGF2		3258		19336		108		1001		231		66		1569		89		76		31		76			
12	X9	A10	60	Wnt3a		2331		16840		96		804		275		59		1339		94		61		18		75			
13	X10	B10	30	Wnt3a		2552		18177		83		991		333		51		1493		105		72		33		89			
14	X11	C10	20	Wnt3a		2955		17630		83		836		281		55		1484		77		73		24		80			
15	X12	D10	10	Wnt3a		2165		17796		101		795		308		51		1404		82		67		24		66			
16			???????		???																								
17	X13	E10	60	IGF		5902		16460		427		959		226		125		2522		109		68		27		72			
18	X14	F10	30	IGF		174		2919		400		53		37		58		580		35		43		18		32			
19	X15	G10	20	IGF		5127		19969		1333		1314		452		206		4460		133		66		37		98			
20	X16	H10	10	IGF		4912		19383		653		1438		421		161		2790		105		73		37		83			
21	X17	A11	60	IFN		3170		17059		173		927		179		76		1596		85		65		26		74			
22	X18	B11	30	IFN		2736		17809		98		765		194		61		1425		79		54		25		81			
23	X19	C11	20	IFN		2956		18120		100		666		149		59		1126		73		54		25		56			
24	X20	D11	10	IFN		2859		17250		127		704		285		59		1373		82		64		17		62			
25	X21	E11	60	BSA		2042		16917		82		736		237		44		1223		78		61		18		66			
26	X22	F11	30	BSA		1601		17511		95		799		272		50		1429		86		60		28		81			
27	X23	G11	20	BSA		2129		17932		87		761		273		47		1253		93		69		22		64			
28	X24	H11	10	BSA		2536		18844		87		813		247		59		1465		84		67		21		81			
29																													
30						p-Akt		p-c-Jun		p-ERK2		p-GSK3a/B		p-JNK		p-MEK1		p-p70 S6 Kinase		p-Stat3 (Ser727)									
31				Wnt oder BSA		2501		17611		91		857		299		54		1430		90	MW								
32						14		3		10		11		9		7		5		14	Std%								
33				Wnt oder BSA		2077		17801		88		777		257		50		1342		85	MW								
34						18		5		6		5		7		13		9		7	Std%								
35				Wnt und BSA		2289		17706		89		817		278		52		1386		87	MW								
36						18		4		8		9		11		11		8		11	Std%								
37						p-Akt		p-c-Jun		p-ERK2		p-GSK3a/B		p-JNK		p-MEK1		p-p70 S6 Kinase		p-Stat3 (Ser727)									
38			60	EGFodIGF		2.3		0.0		0.5		0.6		-0.1		0.7		0.3		0.2									
39			30	EGFodIGF		2.2		0.1		0.6		0.7		0.0		0.4		0.4		0.1									
40			20	EGFodIGF		2.1		0.0		0.4		0.4		-0.3		0.1		0.2		0.1									
41			10	EGFodIGF		1.6		0.1		0.7		0.6		0.5		0.5		0.4		0.3									
42			60	FGFodIFN		0.4		0.0		0.2		0.2		0.1		0.2		0.2		0.0									
43			30	FGFodIFN		0.1		-0.2		0.3		0.2		-0.4		0.2		0.0		0.0									
44			20	FGFodIFN		0.2		0.0		0.1		0.0		-0.3		0.1		0.0		0.0									
45			10	FGFodIFN		0.5		0.1		0.3		0.3		-0.3		0.3		0.2		0.0									
46			60	EGFodIGF		1.4		-0.1		2.3		0.2		-0.3		1.3		0.9		0.3									

mit wnt und bsa normiert, da höchste std 18%

XII | I | III | IV | V | VI | VII | VIII | IX | X | XI | XII | XIII | XIV | XV | Vgl IX X XI XII

Normalansicht | Eingeben | Summe=0



	A	B	C	D	E	F	G	H
1	folder	sample_id			group	condition		
2	Sample_CH_DD_001	1	emp.20 -dox 24h	CH_DD_001	empty empty_cl20	(-)dox	24h	
3	Sample_CH_DD_002	2	emp.20 +dox 24h	CH_DD_002		(+)dox		
4	Sample_CH_DD_003	3	emp.20 -dox 48h	CH_DD_003		(-)dox	48h	
5	Sample_CH_DD_004	4	emp.20 +dox 48h	CH_DD_004		(+)dox		
6	Sample_CH_DD_005	5	emp.20 -dox 72h	CH_DD_005		(-)dox	72h	
7	Sample_CH_DD_006	6	emp.20 +dox 72h	CH_DD_006		(+)dox		
8	Sample_CH_DD_007	7	wt_15 -dox 24h	CH_DD_007	NRas wt_cl15	(-)dox	24h	
9	Sample_CH_DD_008	8	wt_15 +dox 24h	CH_DD_008		(+)dox		
10	Sample_CH_DD_009	9	wt_15 -dox 48h	CH_DD_009		(-)dox	48h	
11	Sample_CH_DD_010	10	wt_15 +dox 48h	CH_DD_010		(+)dox		
12	Sample_CH_DD_011	11	wt_15 -dox 72h	CH_DD_011		(-)dox	72h	
13	Sample_CH_DD_012	12	wt_15 +dox 72h	CH_DD_012		(+)dox		
14	Sample_CH_DD_013	13	12D_6 -dox 24h	CH_DD_013	NRas V12D_cl6	(-)dox	24h	
15	Sample_CH_DD_014	14	12D_6 +dox 24h	CH_DD_014		(+)dox		
16	Sample_CH_DD_015	15	12D_6 -dox 48h	CH_DD_015		(-)dox	48h	
17	Sample_CH_DD_016	16	12D_6 +dox 48h	CH_DD_016		(+)dox		
18	Sample_CH_DD_017	17	12D_6 -dox 72h	CH_DD_017		(-)dox	72h	
19	Sample_CH_DD_018	18	12D_6 +dox 72h	CH_DD_018		(+)dox		
20	Sample_CH_DD_019	19	61K_2 -dox 24h	CH_DD_019	NRas Q61K_cl2	(-)dox	24h	
21	Sample_CH_DD_020	20	61K_2 +dox 24h	CH_DD_020		(+)dox		
22	Sample_CH_DD_021	21	61K_2 -dox 48h	CH_DD_021		(-)dox	48h	
23	Sample_CH_DD_022	22	61K_2 +dox 48h	CH_DD_022		(+)dox		
24	Sample_CH_DD_023	23	61K_2 -dox 72h	CH_DD_023		(-)dox	72h	
25	Sample_CH_DD_024	24	61K_2 +dox 72h	CH_DD_024		(+)dox		
26	Sample_CH_DD_025	25	KRas I -dox 24h	CH_DD_025	KRas G12V_subcl2	(-)dox	24h	
27	Sample_CH_DD_026	26	KRas I +dox 24h	CH_DD_026		(+)dox		
28	Sample_CH_DD_027	27	KRas I -dox 48h	CH_DD_027		(-)dox	48h	
29	Sample_CH_DD_028	28	KRas I +dox 48h	CH_DD_028		(+)dox		
30	Sample_CH_DD_029	29	KRas I -dox 72h	CH_DD_029		(-)dox	72h	
31	Sample_CH_DD_030	30	KRas I +dox 72h	CH_DD_030		(+)dox		
32	Sample_CH_DD_031	31	KRas II -dox 24h	CH_DD_031	KRas G12V_subcl2	(-)dox	24h	
33	Sample_CH_DD_032	32	KRas II +dox 24h	CH_DD_032		(+)dox		
34	Sample_CH_DD_033	33	KRas II -dox484h	CH_DD_033		(-)dox	48h	
35	Sample_CH_DD_034	34	KRas II +dox 48h	CH_DD_034		(+)dox		
36	Sample_CH_DD_035	35	KRas II -dox 72h	CH_DD_035		(-)dox	72h	
37	Sample_CH_DD_036	36	KRas II +dox 72h	CH_DD_036		(+)dox		
38								
39	(+)dox--> oncogene is induced							
40								
41								

	A	B	C	D	E	F	G	H
1	folder	sample_id			group	condition		
2	Sample_CH_DD_001	1	emp.20 -dox 24h	CH_DD_001	empty empty_cl20	(-)dox	24h	
3	Sample_CH_DD_002	2	emp.20 +dox 24h	CH_DD_002		(+)dox		
4	Sample_CH_DD_003	3	emp.20 -dox 48h	CH_DD_003		(-)dox	48h	
5	Sample_CH_DD_004	4	emp.20 +dox 48h	CH_DD_004		(+)dox		
6	Sample_CH_DD_005	5	emp.20 -dox 72h	CH_DD_005		(-)dox	72h	
7	Sample_CH_DD_006	6	emp.20 +dox 72h	CH_DD_006		(+)dox		
8	Sample_CH_DD_007	7	wt_15 -dox 24h	CH_DD_007	NRas wt_cl15	(-)dox	24h	
9	Sample_CH_DD_008	8	wt_15 +dox 24h	CH_DD_008		(+)dox		
10	Sample_CH_DD_009	9	wt_15 -dox 48h	CH_DD_009		(-)dox	48h	
11	Sample_CH_DD_010	10	wt_15 +dox 48h	CH_DD_010		(+)dox		
12	Sample_CH_DD_011	11	wt_15 -dox 72h	CH_DD_011		(-)dox	72h	
13	Sample_CH_DD_012	12	wt_15 +dox 72h	CH_DD_012		(+)dox		
14	Sample_CH_DD_013	13	12D_6 -dox 24h	CH_DD_013	NRas V12D_cl6	(-)dox	24h	
15	Sample_CH_DD_014	14	12D_6 +dox 24h	CH_DD_014		(+)dox		
16	Sample_CH_DD_015	15	12D_6 -dox 48h	CH_DD_015		(-)dox	48h	
17	Sample_CH_DD_016	16	12D_6 +dox 48h	CH_DD_016		(+)dox		
18	Sample_CH_DD_017	17	12D_6 -dox 72h	CH_DD_017		(-)dox	72h	
19	Sample_CH_DD_018	18	12D_6 +dox 72h	CH_DD_018		(+)dox		
20	Sample_CH_DD_019	19	61K_2 -dox 24h	CH_DD_019	NRas Q61K_cl2	(-)dox	24h	
21	Sample_CH_DD_020	20	61K_2 +dox 24h	CH_DD_020		(+)dox		
22	Sample_CH_DD_021	21	61K_2 -dox 48h	CH_DD_021		(-)dox	48h	
23	Sample_CH_DD_022	22	61K_2 +dox 48h	CH_DD_022		(+)dox		
24	Sample_CH_DD_023	23	61K_2 -dox 72h	CH_DD_023		(-)dox	72h	
25	Sample_CH_DD_024	24	61K_2 +dox 72h	CH_DD_024		(+)dox		
26	Sample_CH_DD_025	25	KRas I -dox 24h	CH_DD_025	KRas G12V_subcl2	(-)dox	24h	
27	Sample_CH_DD_026	26	KRas I +dox 24h	CH_DD_026		(+)dox		
28	Sample_CH_DD_027	27	KRas I -dox 48h	CH_DD_027		(-)dox	48h	
29	Sample_CH_DD_028	28	KRas I +dox 48h	CH_DD_028		(+)dox		
30	Sample_CH_DD_029	29	KRas I -dox 72h	CH_DD_029		(-)dox	72h	
31	Sample_CH_DD_030	30	KRas I +dox 72h	CH_DD_030		(+)dox		
32	Sample_CH_DD_031	31	KRas II -dox 24h	CH_DD_031	KRas G12V_subcl2	(-)dox	24h	
33	Sample_CH_DD_032	32	KRas II +dox 24h	CH_DD_032		(+)dox		
34	Sample_CH_DD_033	33	KRas II -dox484h	CH_DD_033		(-)dox	48h	
35	Sample_CH_DD_034	34	KRas II +dox 48h	CH_DD_034		(+)dox		
36	Sample_CH_DD_035	35	KRas II -dox 72h	CH_DD_035		(-)dox	72h	
37	Sample_CH_DD_036	36	KRas II +dox 72h	CH_DD_036		(+)dox		
38								
39	(+)dox--> oncogene is induced							
40								
41								

	A	B	C	D	E	F	G	H
1	folder	sample_id			group	condition		
2	Sample_CH_DD_001	1	emp.20 -dox 24h	CH_DD_001	empty empty_cl20	(-)dox	24h	
3	Sample_CH_DD_002	2	emp.20 +dox 24h	CH_DD_002		(+)dox		
4	Sample_CH_DD_003	3	emp.20 -dox 48h	CH_DD_003		(-)dox	48h	
5	Sample_CH_DD_004	4	emp.20 +dox 48h	CH_DD_004		(+)dox		
6	Sample_CH_DD_005	5	emp.20 -dox 72h	CH_DD_005		(-)dox	72h	
7	Sample_CH_DD_006	6	emp.20 +dox 72h	CH_DD_006		(+)dox		
8	Sample_CH_DD_007	7	wt_15 -dox 24h	CH_DD_007	NRas wt_cl15	(-)dox	24h	
9	Sample_CH_DD_008	8	wt_15 +dox 24h	CH_DD_008		(+)dox		
10	Sample_CH_DD_009	9	wt_15 -dox 48h	CH_DD_009		(-)dox	48h	
11	Sample_CH_DD_010	10	wt_15 +dox 48h	CH_DD_010		(+)dox		
12	Sample_CH_DD_011	11	wt_15 -dox 72h	CH_DD_011		(-)dox	72h	
13	Sample_CH_DD_012	12	wt_15 +dox 72h	CH_DD_012		(+)dox		
14	Sample_CH_DD_013	13	12D_6 -dox 24h	CH_DD_013	NRas V12D_cl6	(-)dox	24h	
15	Sample_CH_DD_014	14	12D_6 +dox 24h	CH_DD_014		(+)dox		
16	Sample_CH_DD_015	15	12D_6 -dox 48h	CH_DD_015		(-)dox	48h	
17	Sample_CH_DD_016	16	12D_6 +dox 48h	CH_DD_016		(+)dox		
18	Sample_CH_DD_017	17	12D_6 -dox 72h	CH_DD_017		(-)dox	72h	
19	Sample_CH_DD_018	18	12D_6 +dox 72h	CH_DD_018		(+)dox		
20	Sample_CH_DD_019	19	61K_2 -dox 24h	CH_DD_019	NRas Q61K_cl2	(-)dox	24h	
21	Sample_CH_DD_020	20	61K_2 +dox 24h	CH_DD_020		(+)dox		
22	Sample_CH_DD_021	21	61K_2 -dox 48h	CH_DD_021		(-)dox	48h	
23	Sample_CH_DD_022	22	61K_2 +dox 48h	CH_DD_022		(+)dox		
24	Sample_CH_DD_023	23	61K_2 -dox 72h	CH_DD_023		(-)dox	72h	
25	Sample_CH_DD_024	24	61K_2 +dox 72h	CH_DD_024		(+)dox		
26	Sample_CH_DD_025	25	KRas I -dox 24h	CH_DD_025	KRas G12V_subcl2	(-)dox	24h	
27	Sample_CH_DD_026	26	KRas I +dox 24h	CH_DD_026		(+)dox		
28	Sample_CH_DD_027	27	KRas I -dox 48h	CH_DD_027		(-)dox	48h	
29	Sample_CH_DD_028	28	KRas I +dox 48h	CH_DD_028		(+)dox		
30	Sample_CH_DD_029	29	KRas I -dox 72h	CH_DD_029		(-)dox	72h	
31	Sample_CH_DD_030	30	KRas I +dox 72h	CH_DD_030		(+)dox		
32	Sample_CH_DD_031	31	KRas II -dox 24h	CH_DD_031	KRas G12V_subcl2	(-)dox	24h	
33	Sample_CH_DD_032	32	KRas II +dox 24h	CH_DD_032		(+)dox		
34	Sample_CH_DD_033	33	KRas II -dox484h	CH_DD_033		(-)dox	48h	
35	Sample_CH_DD_034	34	KRas II +dox 48h	CH_DD_034		(+)dox		
36	Sample_CH_DD_035	35	KRas II -dox 72h	CH_DD_035		(-)dox	72h	
37	Sample_CH_DD_036	36	KRas II +dox 72h	CH_DD_036		(+)dox		
38								
39	(+)dox--> oncogene is induced							
40								
41								

	A	B	C	D	E	F	G	H
1	folder	sample_id			group	condition		
2	Sample_CH_DD_001	1	emp.20 -dox 24h	CH_DD_001	empty empty_cl20	(-)dox	24h	
3	Sample_CH_DD_002	2	emp.20 +dox 24h	CH_DD_002		(+)dox		
4	Sample_CH_DD_003	3	emp.20 -dox 48h	CH_DD_003		(-)dox	48h	
5	Sample_CH_DD_004	4	emp.20 +dox 48h	CH_DD_004		(+)dox		
6	Sample_CH_DD_005	5	emp.20 -dox 72h	CH_DD_005		(-)dox	72h	
7	Sample_CH_DD_006	6	emp.20 +dox 72h	CH_DD_006		(+)dox		
8	Sample_CH_DD_007	7	wt_15 -dox 24h	CH_DD_007	NRas wt_cl15	(-)dox	24h	
9	Sample_CH_DD_008	8	wt_15 +dox 24h	CH_DD_008		(+)dox		
10	Sample_CH_DD_009	9	wt_15 -dox 48h	CH_DD_009		(-)dox	48h	
11	Sample_CH_DD_010	10	wt_15 +dox 48h	CH_DD_010		(+)dox		
12	Sample_CH_DD_011	11	wt_15 -dox 72h	CH_DD_011		(-)dox	72h	
13	Sample_CH_DD_012	12	wt_15 +dox 72h	CH_DD_012		(+)dox		
14	Sample_CH_DD_013	13	12D_6 -dox 24h	CH_DD_013	NRas V12D_cl6	(-)dox	24h	
15	Sample_CH_DD_014	14	12D_6 +dox 24h	CH_DD_014		(+)dox		
16	Sample_CH_DD_015	15	12D_6 -dox 48h	CH_DD_015		(-)dox	48h	
17	Sample_CH_DD_016	16	12D_6 +dox 48h	CH_DD_016		(+)dox		
18	Sample_CH_DD_017	17	12D_6 -dox 72h	CH_DD_017		(-)dox	72h	
19	Sample_CH_DD_018	18	12D_6 +dox 72h	CH_DD_018		(+)dox		
20	Sample_CH_DD_019	19	61K_2 -dox 24h	CH_DD_019	NRas Q61K_cl2	(-)dox	24h	
21	Sample_CH_DD_020	20	61K_2 +dox 24h	CH_DD_020		(+)dox		
22	Sample_CH_DD_021	21	61K_2 -dox 48h	CH_DD_021		(-)dox	48h	
23	Sample_CH_DD_022	22	61K_2 +dox 48h	CH_DD_022		(+)dox		
24	Sample_CH_DD_023	23	61K_2 -dox 72h	CH_DD_023		(-)dox	72h	
25	Sample_CH_DD_024	24	61K_2 +dox 72h	CH_DD_024		(+)dox		
26	Sample_CH_DD_025	25	KRas I -dox 24h	CH_DD_025	KRas G12V_subcl2	(-)dox	24h	
27	Sample_CH_DD_026	26	KRas I +dox 24h	CH_DD_026		(+)dox		
28	Sample_CH_DD_027	27	KRas I -dox 48h	CH_DD_027		(-)dox	48h	
29	Sample_CH_DD_028	28	KRas I +dox 48h	CH_DD_028		(+)dox		
30	Sample_CH_DD_029	29	KRas I -dox 72h	CH_DD_029		(-)dox	72h	
31	Sample_CH_DD_030	30	KRas I +dox 72h	CH_DD_030		(+)dox		
32	Sample_CH_DD_031	31	KRas II -dox 24h	CH_DD_031	KRas G12V_subcl2	(-)dox	24h	
33	Sample_CH_DD_032	32	KRas II +dox 24h	CH_DD_032		(+)dox		
34	Sample_CH_DD_033	33	KRas II -dox484h	CH_DD_033		(-)dox	48h	
35	Sample_CH_DD_034	34	KRas II +dox 48h	CH_DD_034		(+)dox		
36	Sample_CH_DD_035	35	KRas II -dox 72h	CH_DD_035		(-)dox	72h	
37	Sample_CH_DD_036	36	KRas II +dox 72h	CH_DD_036		(+)dox		
38								
39	(+)dox--> oncogene is induced							
40								
41								

	A	B	C	D	E	F	G	H
1	folder	sample_id			group	condition		
2	Sample_CH_DD_001	1	emp.20 -dox 24h	CH_DD_001	empty empty_cl20	(-)dox	24h	
3	Sample_CH_DD_002	2	emp.20 +dox 24h	CH_DD_002		(+)dox		
4	Sample_CH_DD_003	3	emp.20 -dox 48h	CH_DD_003		(-)dox	48h	
5	Sample_CH_DD_004	4	emp.20 +dox 48h	CH_DD_004		(+)dox		
6	Sample_CH_DD_005	5	emp.20 -dox 72h	CH_DD_005		(-)dox	72h	
7	Sample_CH_DD_006	6	emp.20 +dox 72h	CH_DD_006		(+)dox		
8	Sample_CH_DD_007	7	wt_15 -dox 24h	CH_DD_007	NRas wt_cl15	(-)dox	24h	
9	Sample_CH_DD_008	8	wt_15 +dox 24h	CH_DD_008		(+)dox		
10	Sample_CH_DD_009	9	wt_15 -dox 48h	CH_DD_009		(-)dox	48h	
11	Sample_CH_DD_010	10	wt_15 +dox 48h	CH_DD_010		(+)dox		
12	Sample_CH_DD_011	11	wt_15 -dox 72h	CH_DD_011		(-)dox	72h	
13	Sample_CH_DD_012	12	wt_15 +dox 72h	CH_DD_012		(+)dox		
14	Sample_CH_DD_013	13	12D_6 -dox 24h	CH_DD_013	NRas V12D_cl6	(-)dox	24h	
15	Sample_CH_DD_014	14	12D_6 +dox 24h	CH_DD_014		(+)dox		
16	Sample_CH_DD_015	15	12D_6 -dox 48h	CH_DD_015		(-)dox	48h	
17	Sample_CH_DD_016	16	12D_6 +dox 48h	CH_DD_016		(+)dox		
18	Sample_CH_DD_017	17	12D_6 -dox 72h	CH_DD_017		(-)dox	72h	
19	Sample_CH_DD_018	18	12D_6 +dox 72h	CH_DD_018		(+)dox		
20	Sample_CH_DD_019	19	61K_2 -dox 24h	CH_DD_019	NRas Q61K_cl2	(-)dox	24h	
21	Sample_CH_DD_020	20	61K_2 +dox 24h	CH_DD_020		(+)dox		
22	Sample_CH_DD_021	21	61K_2 -dox 48h	CH_DD_021		(-)dox	48h	
23	Sample_CH_DD_022	22	61K_2 +dox 48h	CH_DD_022		(+)dox		
24	Sample_CH_DD_023	23	61K_2 -dox 72h	CH_DD_023		(-)dox	72h	
25	Sample_CH_DD_024	24	61K_2 +dox 72h	CH_DD_024		(+)dox		
26	Sample_CH_DD_025	25	KRas I -dox 24h	CH_DD_025	KRas G12V_subcl2	(-)dox	24h	
27	Sample_CH_DD_026	26	KRas I +dox 24h	CH_DD_026		(+)dox		
28	Sample_CH_DD_027	27	KRas I -dox 48h	CH_DD_027		(-)dox	48h	
29	Sample_CH_DD_028	28	KRas I +dox 48h	CH_DD_028		(+)dox		
30	Sample_CH_DD_029	29	KRas I -dox 72h	CH_DD_029		(-)dox	72h	
31	Sample_CH_DD_030	30	KRas I +dox 72h	CH_DD_030		(+)dox		
32	Sample_CH_DD_031	31	KRas II -dox 24h	CH_DD_031	KRas G12V_subcl2	(-)dox	24h	
33	Sample_CH_DD_032	32	KRas II +dox 24h	CH_DD_032		(+)dox		
34	Sample_CH_DD_033	33	KRas II -dox484h	CH_DD_033		(-)dox	48h	
35	Sample_CH_DD_034	34	KRas II +dox 48h	CH_DD_034		(+)dox		
36	Sample_CH_DD_035	35	KRas II -dox 72h	CH_DD_035		(-)dox	72h	
37	Sample_CH_DD_036	36	KRas II +dox 72h	CH_DD_036		(+)dox		
38								
39	(+)dox--> oncogene is induced							
40								
41								

```
read_excel("untidy_data.xlsx") %>%
  set_colnames(mynames) %>%
  slice(1:36) %>%
  fill(group, condition) %>%
  separate(group, into = c("Gene", "Mutation", "clone"), sep = "_") %>%
  write_tsv("tidy_data.tsv")
```

Go to file/function | Addins | OncoPATH_RAS

Filter

	Sample_ID	Gene	Mutation	Clone	dox	time	replicate	Sample_Name	dox_logical
1	CH_DD_001	empty	empty	cl20	(-)	24h	1	emp_20 -dox 24h	FALSE
2	CH_DD_002	empty	empty	cl20	(+)	24h	1	emp_20 +dox 24h	TRUE
3	CH_DD_003	empty	empty	cl20	(-)	48h	1	emp_20 -dox 48h	FALSE
4	CH_DD_004	empty	empty	cl20	(+)	48h	1	emp_20 +dox 48h	TRUE
5	CH_DD_005	empty	empty	cl20	(-)	72h	1	emp_20 -dox 72h	FALSE
6	CH_DD_006	empty	empty	cl20	(+)	72h	1	emp_20 +dox 72h	TRUE
7	CH_DD_007	NRas	wt	cl15	(-)	24h	1	wt_15 -dox 24h	FALSE
8	CH_DD_008	NRas	wt	cl15	(+)	24h	1	wt_15 +dox 24h	TRUE
9	CH_DD_009	NRas	wt	cl15	(-)	48h	1	wt_15 -dox 48h	FALSE
10	CH_DD_010	NRas	wt	cl15	(+)	48h	1	wt_15 +dox 48h	TRUE
11	CH_DD_011	NRas	wt	cl15	(-)	72h	1	wt_15 -dox 72h	FALSE
12	CH_DD_012	NRas	wt	cl15	(+)	72h	1	wt_15 +dox 72h	TRUE
13	CH_DD_013	NRas	G12D	cl6	(-)	24h	1	12D_6 -dox 24h	FALSE
14	CH_DD_014	NRas	G12D	cl6	(+)	24h	1	12D_6 +dox 24h	TRUE
15	CH_DD_015	NRas	G12D	cl6	(-)	48h	1	12D_6 -dox 48h	FALSE
16	CH_DD_016	NRas	G12D	cl6	(+)	48h	1	12D_6 +dox 48h	TRUE
17	CH_DD_017	NRas	G12D	cl6	(-)	72h	1	12D_6 -dox 72h	FALSE
18	CH_DD_018	NRas	G12D	cl6	(+)	72h	1	12D_6 +dox 72h	TRUE

Showing 1 to 19 of 36 entries

```
read_excel("untidy_data.xlsx") %>%
  set_colnames(mynames) %>%
  slice(1:36) %>%
  fill(group, condition) %>%
  separate(group, into = c("Gene", "Mutation", "clone"), sep = "_") %>%
  write_tsv("tidy_data.tsv")
```

`read_excel`

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx")
Source: local data frame [38 x 7]

            folder sample_id              NA       NA             group condition    NA
             (chr)     (dbl)           (chr)    (chr)             (chr)     (chr) (chr)
1  Sample_CH_DD_001         1 emp.20 -dox 24h CH_DD_001 empty empty_cl20   (-)dox    24h
2  Sample_CH_DD_002         2 emp.20 +dox 24h CH_DD_002                NA   (+)dox     NA
3  Sample_CH_DD_003         3 emp.20 -dox 48h CH_DD_003                NA   (-)dox    48h
4  Sample_CH_DD_004         4 emp.20 +dox 48h CH_DD_004                NA   (+)dox     NA
5  Sample_CH_DD_005         5 emp.20 -dox 72h CH_DD_005                NA   (-)dox    72h
6  Sample_CH_DD_006         6 emp.20 +dox 72h CH_DD_006                NA   (+)dox     NA
7  Sample_CH_DD_007         7  wt_15 -dox 24h CH_DD_007     NRas wt_cl15   (-)dox    24h
8  Sample_CH_DD_008         8  wt_15 +dox 24h CH_DD_008                NA   (+)dox     NA
9  Sample_CH_DD_009         9  wt_15 -dox 48h CH_DD_009                NA   (-)dox    48h
10 Sample_CH_DD_010        10  wt_15 +dox 48h CH_DD_010                NA   (+)dox     NA
..             ...       ...             ...      ...               ...       ...   ...
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7))
Source: local data frame [38 x 7]

                X1    X2              X3       X4                  X5      X6    X7
             (chr) (dbl)           (chr)    (chr)               (chr)   (chr) (chr)
1  Sample_CH_DD_001     1 emp.20 -dox 24h CH_DD_001 empty empty_cl20 (-)dox    24h
2  Sample_CH_DD_002     2 emp.20 +dox 24h CH_DD_002                NA (+)dox     NA
3  Sample_CH_DD_003     3 emp.20 -dox 48h CH_DD_003                NA (-)dox    48h
4  Sample_CH_DD_004     4 emp.20 +dox 48h CH_DD_004                NA (+)dox     NA
5  Sample_CH_DD_005     5 emp.20 -dox 72h CH_DD_005                NA (-)dox    72h
6  Sample_CH_DD_006     6 emp.20 +dox 72h CH_DD_006                NA (+)dox     NA
7  Sample_CH_DD_007     7  wt_15 -dox 24h CH_DD_007     NRas wt_cl15 (-)dox    24h
8  Sample_CH_DD_008     8  wt_15 +dox 24h CH_DD_008                NA (+)dox     NA
9  Sample_CH_DD_009     9  wt_15 -dox 48h CH_DD_009                NA (-)dox    48h
10 Sample_CH_DD_010    10  wt_15 +dox 48h CH_DD_010                NA (+)dox     NA
..              ...   ...             ...      ...                 ...     ...   ...
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% tail
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     tail(10)
Source: local data frame [10 x 7]

                          X1    X2              X3       X4                 X5      X6    X7
                       (chr) (dbl)           (chr)    (chr)              (chr)   (chr) (chr)
1              Sample_CH_DD_029    29  KRas I -dox 72h CH_DD_029                NA (-)dox    72h
2              Sample_CH_DD_030    30  KRas I +dox 72h CH_DD_030                NA (+)dox     NA
3              Sample_CH_DD_031    31 KRas II -dox 24h CH_DD_031 KRas G12V_subcl2 (-)dox    24h
4              Sample_CH_DD_032    32 KRas II +dox 24h CH_DD_032                NA (+)dox     NA
5              Sample_CH_DD_033    33 KRas II -dox484h CH_DD_033                NA (-)dox    48h
6              Sample_CH_DD_034    34 KRas II +dox 48h CH_DD_034                NA (+)dox     NA
7              Sample_CH_DD_035    35 KRas II -dox 72h CH_DD_035                NA (-)dox    72h
8              Sample_CH_DD_036    36 KRas II +dox 72h CH_DD_036                NA (+)dox     NA
9  (+)dox--> oncogene is induced    NA              NA       NA                 NA      NA    NA
10                            NA    NA              NA       NA                 NA      NA    NA
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7))
Source: local data frame [38 x 7]

                X1    X2              X3       X4                  X5      X6    X7
             (chr) (dbl)           (chr)    (chr)               (chr)   (chr) (chr)
1  Sample_CH_DD_001     1 emp.20 -dox 24h CH_DD_001 empty empty_cl20 (-)dox    24h
2  Sample_CH_DD_002     2 emp.20 +dox 24h CH_DD_002                NA (+)dox     NA
3  Sample_CH_DD_003     3 emp.20 -dox 48h CH_DD_003                NA (-)dox    48h
4  Sample_CH_DD_004     4 emp.20 +dox 48h CH_DD_004                NA (+)dox     NA
5  Sample_CH_DD_005     5 emp.20 -dox 72h CH_DD_005                NA (-)dox    72h
6  Sample_CH_DD_006     6 emp.20 +dox 72h CH_DD_006                NA (+)dox     NA
7  Sample_CH_DD_007     7  wt_15 -dox 24h CH_DD_007     NRas wt_cl15 (-)dox    24h
8  Sample_CH_DD_008     8  wt_15 +dox 24h CH_DD_008                NA (+)dox     NA
9  Sample_CH_DD_009     9  wt_15 -dox 48h CH_DD_009                NA (-)dox    48h
10 Sample_CH_DD_010    10  wt_15 +dox 48h CH_DD_010                NA (+)dox     NA
..              ...   ...             ...      ...                 ...     ...   ...
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% slice
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     slice(1:36)
Source: local data frame [36 x 7]

                X1    X2              X3       X4                 X5      X6    X7
             (chr) (dbl)           (chr)    (chr)              (chr)   (chr) (chr)
1  Sample_CH_DD_001     1 emp.20 -dox 24h CH_DD_001 empty empty_cl20 (-)dox    24h
2  Sample_CH_DD_002     2 emp.20 +dox 24h CH_DD_002                NA (+)dox     NA
3  Sample_CH_DD_003     3 emp.20 -dox 48h CH_DD_003                NA (-)dox    48h
4  Sample_CH_DD_004     4 emp.20 +dox 48h CH_DD_004                NA (+)dox     NA
5  Sample_CH_DD_005     5 emp.20 -dox 72h CH_DD_005                NA (-)dox    72h
6  Sample_CH_DD_006     6 emp.20 +dox 72h CH_DD_006                NA (+)dox     NA
7  Sample_CH_DD_007     7  wt_15 -dox 24h CH_DD_007    NRas wt_cl15 (-)dox    24h
8  Sample_CH_DD_008     8  wt_15 +dox 24h CH_DD_008                NA (+)dox     NA
9  Sample_CH_DD_009     9  wt_15 -dox 48h CH_DD_009                NA (-)dox    48h
10 Sample_CH_DD_010    10  wt_15 +dox 48h CH_DD_010                NA (+)dox     NA
..              ...   ...             ...      ...                ...     ...   ...
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% slice %>% fill
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     slice(1:36) %>%
+     fill(X5:X7)
Source: local data frame [36 x 7]

                X1    X2              X3       X4                 X5      X6    X7
             (chr) (dbl)           (chr)    (chr)              (chr)   (chr) (chr)
1  Sample_CH_DD_001     1 emp.20 -dox 24h CH_DD_001 empty empty_cl20 (-)dox    24h
2  Sample_CH_DD_002     2 emp.20 +dox 24h CH_DD_002 empty empty_cl20 (+)dox    24h
3  Sample_CH_DD_003     3 emp.20 -dox 48h CH_DD_003 empty empty_cl20 (-)dox    48h
4  Sample_CH_DD_004     4 emp.20 +dox 48h CH_DD_004 empty empty_cl20 (+)dox    48h
5  Sample_CH_DD_005     5 emp.20 -dox 72h CH_DD_005 empty empty_cl20 (-)dox    72h
6  Sample_CH_DD_006     6 emp.20 +dox 72h CH_DD_006 empty empty_cl20 (+)dox    72h
7  Sample_CH_DD_007     7  wt_15 -dox 24h CH_DD_007     NRas wt_cl15 (-)dox    24h
8  Sample_CH_DD_008     8  wt_15 +dox 24h CH_DD_008     NRas wt_cl15 (+)dox    24h
9  Sample_CH_DD_009     9  wt_15 -dox 48h CH_DD_009     NRas wt_cl15 (-)dox    48h
10 Sample_CH_DD_010    10  wt_15 +dox 48h CH_DD_010     NRas wt_cl15 (+)dox    48h
..              ...   ...           ...      ...                ...     ...   ...
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% slice %>% fill %>% select
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     slice(1:36) %>%
+     fill(X5:X7) %>%
+     select(X1, X5)
Source: local data frame [36 x 2]

                X1              X5
             (chr)           (chr)
1  Sample_CH_DD_001 empty empty_cl20
2  Sample_CH_DD_002 empty empty_cl20
3  Sample_CH_DD_003 empty empty_cl20
4  Sample_CH_DD_004 empty empty_cl20
5  Sample_CH_DD_005 empty empty_cl20
6  Sample_CH_DD_006 empty empty_cl20
7  Sample_CH_DD_007     NRas wt_cl15
8  Sample_CH_DD_008     NRas wt_cl15
9  Sample_CH_DD_009     NRas wt_cl15
10 Sample_CH_DD_010     NRas wt_cl15
..             ...             ...
> |
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% slice %>% fill %>% select %>% distinct
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
> 
> 
> 
> 
> 
> 
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     slice(1:36) %>%
+     fill(X5:X7) %>%
+     select(X1, X5) %>%
+     distinct(X5)
Source: local data frame [5 x 2]

                X1                X5
             (chr)             (chr)
1 Sample_CH_DD_001 empty empty_cl20
2 Sample_CH_DD_007     NRas wt_cl15
3 Sample_CH_DD_013    NRas V12D_cl6
4 Sample_CH_DD_019    NRas Q61K_cl2
5 Sample_CH_DD_025 KRas G12V_subcl2
> 
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% slice %>% fill %>% select %>% distinct %>%
    separate
```

Caution!
`readr, tidy & dplyr` do “clever” stuff.
(heuristics like predicting a column class by looking at the first 1000 entries)

Console ~/Projects/R/Current/dplyr_tut/

```
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     slice(1:36) %>%
+     fill(X5:X7) %>%
+     select(X1, X5) %>%
+     distinct(X5) %>%
+     separate(X5, into = c("Gene", "Mutation", "Clone"))
Source: local data frame [5 x 4]

                X1  Gene Mutation  Clone
             (chr) (chr)    (chr)  (chr)
1 Sample_CH_DD_001 empty    empty   cl20
2 Sample_CH_DD_007  NRas       wt   cl15
3 Sample_CH_DD_013  NRas     V12D    cl6
4 Sample_CH_DD_019  NRas     Q61K    cl2
5 Sample_CH_DD_025  KRas     G12V subcl2
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% slice %>% fill %>% select %>% distinct
    separate
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     slice(1:36) %>%
+     fill(X5:X7) %>%
+     select(X1, X5) %>%
+     distinct(X5) %>%
+     separate(X5, into = c("Gene", "Mutation", "Clone"), sep = " |_")
Source: local data frame [5 x 4]

                X1  Gene Mutation  Clone
             (chr) (chr)    (chr)  (chr)
1 Sample_CH_DD_001 empty    empty   cl20
2 Sample_CH_DD_007  NRas       wt   cl15
3 Sample_CH_DD_013  NRas     V12D    cl6
4 Sample_CH_DD_019  NRas     Q61K    cl2
5 Sample_CH_DD_025  KRas     G12V subcl2
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% slice %>% fill %>% select %>% distinct
    separate %>% unite
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     slice(1:36) %>%
+     fill(X5:X7) %>%
+     select(X1, X5) %>%
+     distinct(X5) %>%
+     separate(X5, into = c("Gene", "Mutation", "Clone"), sep = " |_") %>%
+     unite(G_M_C, Gene, Mutation, Clone)
Source: local data frame [5 x 2]

                X1           G_M_C
             (chr)           (chr)
1 Sample_CH_DD_001 empty_empty_cl20
2 Sample_CH_DD_007     NRas_wt_cl15
3 Sample_CH_DD_013    NRas_V12D_cl6
4 Sample_CH_DD_019    NRas_Q61K_cl2
5 Sample_CH_DD_025 KRas_G12V_subcl2
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

```
read_excel %>% set_colnames %>% slice %>% fill %>% select %>% distinct
    separate %>% unite
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
>
> read_excel("annotation_rnaseq_collapsed.xlsx") %>%
+     set_colnames(paste0("X",1:7)) %>%
+     slice(1:36) %>%
+     fill(X5:X7) %>%
+     select(X1, X5) %>%
+     distinct(X5) %>%
+     separate(X5, into = c("Gene", "Mutation", "Clone"), sep = " |_") %>%
+     unite(G_M_C, Gene, Mutation, Clone, sep = "  (-_-)  ")
Source: local data frame [5 x 2]

                X1                          G_M_C
             (chr)                          (chr)
1 Sample_CH_DD_001 empty  (-_-)  empty  (-_-)  cl20
2 Sample_CH_DD_007     NRas  (-_-)  wt  (-_-)  cl15
3 Sample_CH_DD_013    NRas  (-_-)  V12D  (-_-)  cl6
4 Sample_CH_DD_019    NRas  (-_-)  Q61K  (-_-)  cl2
5 Sample_CH_DD_025 KRas  (-_-)  G12V  (-_-)  subcl2
>
```

Environment | Files | Plots | Packages | Help | Git | Viewer

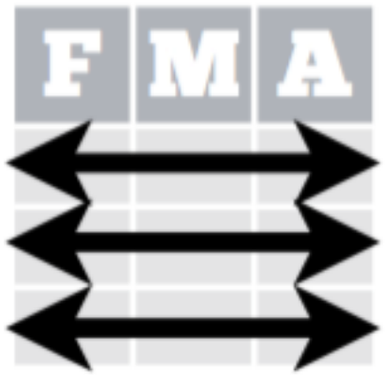
	A	B	C	D	E	F	G	H	I	J	K
1	STUDY_ID	CASE_ID	AKAP9	APC	ARID1A	ATM	ATRX	BRAF	BRCA2	CDKN2A	EGFR
2	acc_tcga	TCGA.OR.A5J1.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	acc_tcga	TCGA.OR.A5J2.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	acc_tcga	TCGA.OR.A5J3.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	acc_tcga	TCGA.OR.A5J4.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	acc_tcga	TCGA.OR.A5J5.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	acc_tcga	TCGA.OR.A5J6.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	acc_tcga	TCGA.OR.A5J7.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	acc_tcga	TCGA.OR.A5J8.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	acc_tcga	TCGA.OR.A5J9.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	acc_tcga	TCGA.OR.A5JA.01	NA	NA	NA	NA	E1365*	NA	NA	NA	NA
12	acc_tcga	TCGA.OR.A5JB.01	NA	NA	NA	NA	NA	D594G	NA	NA	NA
13	acc_tcga	TCGA.OR.A5JC.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	acc_tcga	TCGA.OR.A5JD.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	acc_tcga	TCGA.OR.A5JE.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
16	acc_tcga	TCGA.OR.A5JF.01	NA	NA	NA	NA	N1754K	NA	NA	NA	NA
17	acc_tcga	TCGA.OR.A5JG.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	acc_tcga	TCGA.OR.A5JH.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	acc_tcga	TCGA.OR.A5JI.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	acc_tcga	TCGA.OR.A5JJ.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
21	acc_tcga	TCGA.OR.A5JK.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	acc_tcga	TCGA.OR.A5JL.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
23	acc_tcga	TCGA.OR.A5JM.01	NA	NA	NA	X633_splice	NA	NA	NA	A30Gfs*18	NA
24	acc_tcga	TCGA.OR.A5JO.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
25	acc_tcga	TCGA.OR.A5JP.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
26	acc_tcga	TCGA.OR.A5JQ.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
27	acc_tcga	TCGA.OR.A5JR.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
28	acc_tcga	TCGA.OR.A5JS.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
29	acc_tcga	TCGA.OR.A5JT.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
30	acc_tcga	TCGA.OR.A5JU.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
31	acc_tcga	TCGA.OR.A5JV.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
32	acc_tcga	TCGA.OR.A5JW.01	L1273I	NA	NA	NA	NA	NA	NA	NA	NA
33	acc_tcga	TCGA.OR.A5JX.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
34	acc_tcga	TCGA.OR.A5JY.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
35	acc_tcga	TCGA.OR.A5JZ.01	NA	NA	NA	NA	NA	NA	NA	NA	NA
36	acc_tcga	TCGA.OR.A5K0.01	NA	S1465Wfs*3	NA	NA	K467*	NA	NA	NA	R377K

Tidy data definition

In a tidy data set:

Each **variable** is saved in its own **column**

&



Each **observation** is saved in its own **row**

Wickham, H. (2014). Tidy Data. Journal of Statistical Software

`read_tsv`

Go to file/function

dplyr_tut

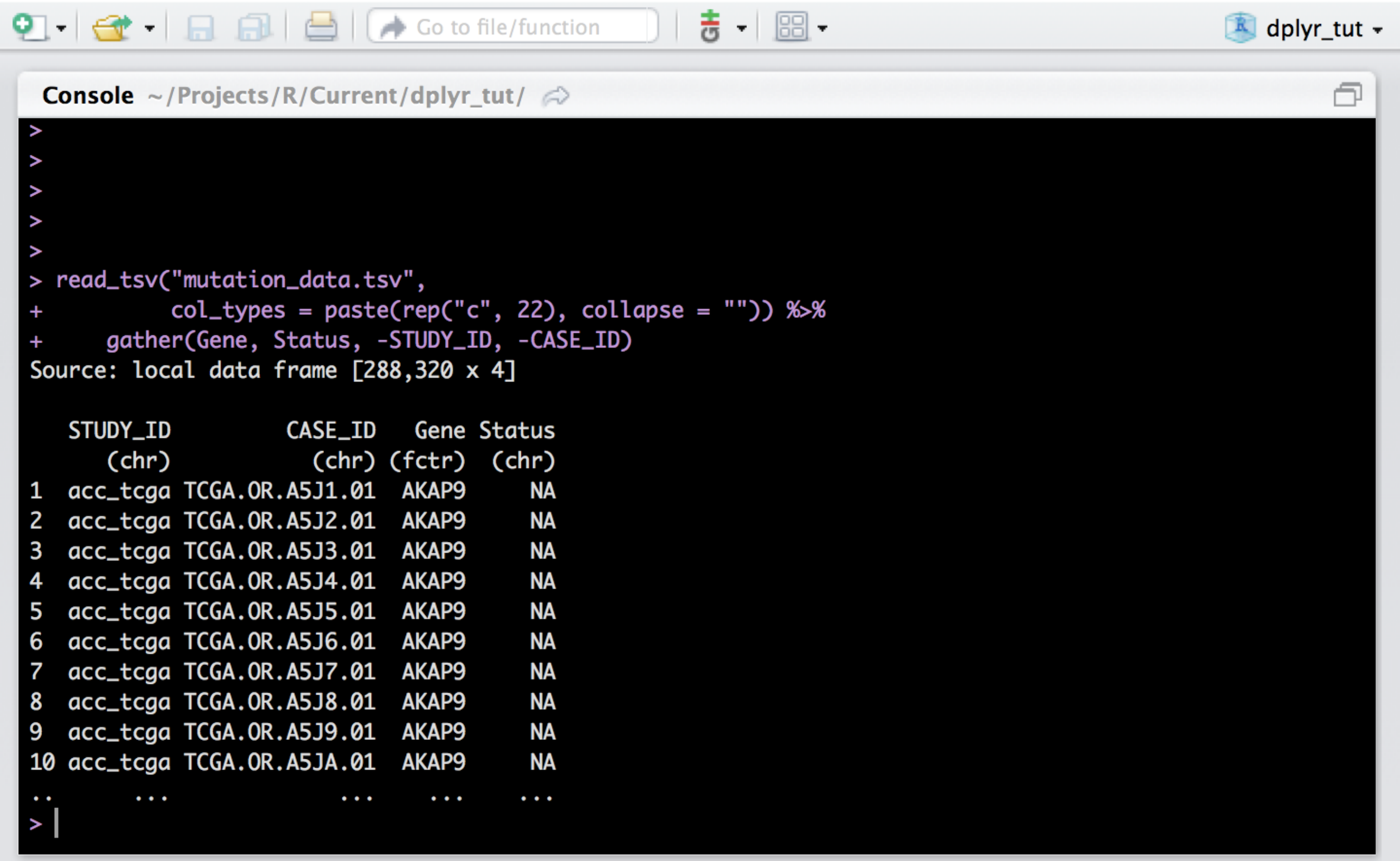
Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
>
> read_tsv("mutation_data.tsv",
+          col_types = paste(rep("c", 22), collapse = ""))
Source: local data frame [14,416 x 22]
```

	STUDY_ID (chr)	CASE_ID (chr)	AKAP9 (chr)	APC (chr)	ARID1A (chr)	ATM (chr)	ATRX (chr)	BRAF (chr)	BRCA2 (chr)	CDKN2A (chr)	EGFR (chr)	KMT2C (chr)	KMT2D (chr)
1	acc_tcga	TCGA.OR.A5J1.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	acc_tcga	TCGA.OR.A5J2.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	acc_tcga	TCGA.OR.A5J3.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	acc_tcga	TCGA.OR.A5J4.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	R894Q	P754A
5	acc_tcga	TCGA.OR.A5J5.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	P2354Lfs*30
6	acc_tcga	TCGA.OR.A5J6.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	acc_tcga	TCGA.OR.A5J7.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	acc_tcga	TCGA.OR.A5J8.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	acc_tcga	TCGA.OR.A5J9.01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	acc_tcga	TCGA.OR.A5JA.01	NA	NA	NA	NA	E1365*	NA	NA	NA	NA	E1828*	NA
..	...	...	...	...	...	...	...	...	...	...	...	...	...

```
Variables not shown: KRAS (chr), NF1 (chr), PBRM1 (chr), PIK3CA (chr), PTEN (chr), SETD2 (chr), SPEN
  (chr), TP53 (chr), VHL (chr)
>
```

```
read_tsv %>% gather(key, value, -variable)
```



```
read_tsv %>% gather %>% spread(key, value)
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
> read_tsv("mutation_data.tsv",
+          col_types = paste(rep("c", 22), collapse = "")) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID) %>%
+     spread(Gene, Status)
Source: local data frame [14,416 x 22]

   STUDY_ID         CASE_ID AKAP9   APC ARID1A   ATM   ATRX  BRAF BRCA2 CDKN2A  EGFR  KMT2C       KMT2D
      (chr)           (chr) (chr) (chr)  (chr) (chr)  (chr) (chr) (chr)  (chr) (chr)  (chr)       (chr)
1  acc_tcga TCGA.OR.A5J1.01    NA    NA     NA    NA     NA    NA    NA     NA    NA     NA          NA
2  acc_tcga TCGA.OR.A5J2.01    NA    NA     NA    NA     NA    NA    NA     NA    NA     NA          NA
3  acc_tcga TCGA.OR.A5J3.01    NA    NA     NA    NA     NA    NA    NA     NA    NA     NA          NA
4  acc_tcga TCGA.OR.A5J4.01    NA    NA     NA    NA     NA    NA    NA     NA    NA  R894Q       P754A
5  acc_tcga TCGA.OR.A5J5.01    NA    NA     NA    NA     NA    NA    NA     NA    NA     NA P2354Lfs*30
6  acc_tcga TCGA.OR.A5J6.01    NA    NA     NA    NA     NA    NA    NA     NA    NA     NA          NA
7  acc_tcga TCGA.OR.A5J7.01    NA    NA     NA    NA     NA    NA    NA     NA    NA     NA          NA
8  acc_tcga TCGA.OR.A5J8.01    NA    NA     NA    NA     NA    NA    NA     NA    NA     NA          NA
9  acc_tcga TCGA.OR.A5J9.01    NA    NA     NA    NA     NA    NA    NA     NA    NA     NA          NA
10 acc_tcga TCGA.OR.A5JA.01    NA    NA     NA    NA E1365*    NA    NA     NA    NA E1828*          NA
..      ...             ...   ...   ...    ...   ...    ...   ...   ...    ...   ...    ...         ...
Variables not shown: KRAS (chr), NF1 (chr), PBRM1 (chr), PIK3CA (chr), PTEN (chr), SETD2 (chr), SPEN
  (chr), TP53 (chr), VHL (chr)
>
```

```
read_tsv %>% gather
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
>
> read_tsv("mutation_data.tsv",
+          col_types = paste(rep("c", 22), collapse = "")) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID) %>%
+     spread(Gene, Status) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID, -BRAF)
Source: local data frame [273,904 x 5]

   STUDY_ID         CASE_ID  BRAF   Gene Status
      (chr)           (chr) (chr) (fctr)  (chr)
1  acc_tcga TCGA.OR.A5J1.01    NA  AKAP9     NA
2  acc_tcga TCGA.OR.A5J2.01    NA  AKAP9     NA
3  acc_tcga TCGA.OR.A5J3.01    NA  AKAP9     NA
4  acc_tcga TCGA.OR.A5J4.01    NA  AKAP9     NA
5  acc_tcga TCGA.OR.A5J5.01    NA  AKAP9     NA
6  acc_tcga TCGA.OR.A5J6.01    NA  AKAP9     NA
7  acc_tcga TCGA.OR.A5J7.01    NA  AKAP9     NA
8  acc_tcga TCGA.OR.A5J8.01    NA  AKAP9     NA
9  acc_tcga TCGA.OR.A5J9.01    NA  AKAP9     NA
10 acc_tcga TCGA.OR.A5JA.01    NA  AKAP9     NA
..      ...             ...   ...    ...    ...
>
```

`read_tsv %>% gather %>% filter`

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
>
>
> read_tsv("mutation_data.tsv",
+          col_types = paste(rep("c", 22), collapse = "")) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID) %>%
+     spread(Gene, Status) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID, -BRAF) %>%
+     filter(!is.na(BRAF))
Source: local data frame [15,998 x 5]

       STUDY_ID         CASE_ID  BRAF   Gene        Status
          (chr)           (chr) (chr) (fctr)         (chr)
1       acc_tcga TCGA.OR.A5JB.01 D594G  AKAP9            NA
2      blca_tcga TCGA.GD.A3OP.01 S364L  AKAP9            NA
3  blca_tcga_pub TCGA.GD.A3OP.01 S364L  AKAP9            NA
4      brca_tcga TCGA.AC.A2FF.01 D449N  AKAP9 D3873N,R1947C
5      brca_tcga TCGA.AN.A0FZ.01 K698R  AKAP9            NA
6      brca_tcga TCGA.C8.A1HJ.01 E309*  AKAP9            NA
7      brca_tcga TCGA.E2.A14U.01 E433K  AKAP9            NA
8      brca_tcga TCGA.E2.A15K.06 L537S  AKAP9            NA
9  brca_tcga_pub TCGA.AN.A0FN.01 K698R  AKAP9            NA
10 brca_tcga_pub TCGA.C8.A12P.01 E309*  AKAP9            NA
..           ...             ...   ...    ...           ...
>
```

```
read_tsv %>% gather %>% filter %>% group_by
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
> read_tsv("mutation_data.tsv",
+          col_types = paste(rep("c", 22), collapse = "")) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID) %>%
+     spread(Gene, Status) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID, -BRAF) %>%
+     filter(!is.na(BRAF)) %>%
+     group_by(BRAF)
Source: local data frame [15,998 x 5]
Groups: BRAF [93]

       STUDY_ID         CASE_ID  BRAF   Gene        Status
          (chr)           (chr) (chr) (fctr)         (chr)
1       acc_tcga TCGA.OR.A5JB.01 D594G  AKAP9            NA
2      blca_tcga TCGA.GD.A3OP.01 S364L  AKAP9            NA
3  blca_tcga_pub TCGA.GD.A3OP.01 S364L  AKAP9            NA
4      brca_tcga TCGA.AC.A2FF.01 D449N  AKAP9 D3873N,R1947C
5      brca_tcga TCGA.AN.A0FZ.01 K698R  AKAP9            NA
6      brca_tcga TCGA.C8.A1HJ.01 E309*  AKAP9            NA
7      brca_tcga TCGA.E2.A14U.01 E433K  AKAP9            NA
8      brca_tcga TCGA.E2.A15K.06 L537S  AKAP9            NA
9  brca_tcga_pub TCGA.AN.A0FN.01 K698R  AKAP9            NA
10 brca_tcga_pub TCGA.C8.A12P.01 E309*  AKAP9            NA
..          ...             ...   ...    ...           ...
>
```

```
read_tsv %>% gather %>% filter %>% group_by %>% summarise %>% arrange
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
+           col_types = paste(rep("c", 22), collapse = "")) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID) %>%
+     spread(Gene, Status) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID, -BRAF) %>%
+     filter(!is.na(BRAF)) %>%
+     group_by(BRAF) %>%
+     summarise(n = n()) %>%
+     arrange(-n)
Source: local data frame [93 x 2]

        BRAF     n
       (chr) (int)
1      V600E 12160
2      V600K   380
3     Fusion   171
4      K601E   171
5      N581S   152
6  P403Lfs*8   133
7      G469V   114
8      G466V    95
9      A762V    76
10     G469A    76
..       ...   ...
>
```

```
read_tsv %>% gather %>% filter %>% group_by %>% summarise %>% arrange
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
+     gather(Gene, Status, -STUDY_ID, -CASE_ID) %>%
+     spread(Gene, Status) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID, -BRAF) %>%
+     filter(!is.na(BRAF)) %>%
+     group_by(BRAF, STUDY_ID) %>%
+     summarise(n = n()) %>%
+     ungroup %>%
+     arrange(-n)
Source: local data frame [163 x 3]

     BRAF          STUDY_ID     n
    (chr)             (chr) (int)
1   V600E         thca_tcga  4503
2   V600E     thca_tcga_pub  4465
3   V600E         skcm_tcga  2451
4   V600K         skcm_tcga   380
5   V600E coadread_tcga_pub   361
6  Fusion     thca_tcga_pub   171
7   V600E          gbm_tcga    95
8   V600E  gbm_tcga_pub2013    95
9   V600E         luad_tcga    76
10  V600E     luad_tcga_pub    76
..    ...               ...   ...
>
```

```
read_tsv %>% gather %>% filter %>% group_by %>% summarise %>% arrange
```

Go to file/function

dplyr_tut

Console ~/Projects/R/Current/dplyr_tut/

```
+     gather(Gene, Status, -STUDY_ID, -CASE_ID) %>%
+     spread(Gene, Status) %>%
+     gather(Gene, Status, -STUDY_ID, -CASE_ID, -BRAF) %>%
+     filter(!is.na(BRAF), !is.na(Status)) %>%
+     group_by(Gene, STUDY_ID) %>%
+     summarise(n = n()) %>%
+     ungroup %>%
+     arrange(-n)
Source: local data frame [239 x 3]

     Gene      STUDY_ID     n
   (fctr)         (chr) (int)
1   KMT2D     skcm_tcga    28
2   KMT2C     skcm_tcga    27
3    TP53     skcm_tcga    27
4  CDKN2A     skcm_tcga    25
5    SPEN     skcm_tcga    23
6    PTEN     skcm_tcga    21
7   AKAP9     skcm_tcga    16
8     APC     skcm_tcga    14
9  ARID1A     stad_tcga    13
10 ARID1A stad_tcga_pub    13
..    ...           ...   ...
> 
```

Data Transformation with dplyr : : CHEAT SHEET

dplyr functions work with pipes and expect **tidy data**. In tidy data:

Each **variable** is in its own **column**

&

Each **observation**, or **case**, is in its own **row**

pipes

x %>% f(y) becomes **f(x, y)**

Summarise Cases

These apply **summary functions** to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).

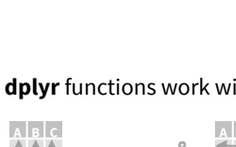
summary function

summarise(.data, …**)**
Compute table of summaries.
summarise(mtcars, avg = mean(mpg))

count(x, ..., wt = NULL, sort = FALSE**)**
Count number of rows in each group defined by the variables in … Also **tally()**.
count(mtcars, cyl)

Group Cases

Use **group_by(**.data, ..., .add = FALSE**)** to create a "grouped" copy of a table grouped by columns in ... dplyr functions will manipulate each "group" separately and combine the results.



```
mtcars %>%
  group_by(cyl) %>%
  summarise(avg = mean(mpg))
```

Use **rowwise(**.data, ...**)** to group data into individual rows. dplyr functions will compute results for each row. Also used to apply functions to list-columns without purrr functions.

```
starwars %>%
  rowwise() %>%
  mutate(film_count = length(films))
```

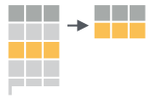
ungroup(x, …**)** Returns ungrouped copy of table.
ungroup(g_mtcars)

RStudio® is a trademark of RStudio, Inc

Manipulate Cases

EXTRACT CASES

Row functions return a subset of rows as a new table.



filter(.data, …**)** Extract rows that meet logical criteria.
filter(mtcars, mpg > 20)

Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.

pull(.data, var = -1**)** Extract column values as a vector. Choose by name or index.
pull(mtcars, wt)

Data Import : : CHEAT SHEET

R's **tidyverse** is built around **tidy data** stored in **tibbles**, which are enhanced data frames.

The front side of this sheet shows how to read text files into R with **readr**.

The reverse side shows how to create tibbles with **tibble** and to layout tidy data with **tidyr**.

OTHER TYPES OF DATA

Try one of the following packages to import other types of files

- **haven** - SPSS, Stata, and SAS files
- **readxl** - excel files (.xls and .xlsx)
- **DBI** - databases
- **jsonlite** - json
- **xml2** - XML
- **httr** - Web APIs
- **rvest** - HTML (Web Scraping)

Save Data

Save **x**, an R object, to **path**, a file path, as:

Comma delimited file
write_csv(x, path, na = "NA", append = FALSE, col_names = !append**)**

File with arbitrary delimiter
write_delim(x, path, delim = " ", na = "NA", append = FALSE, col_names = !append**)**

CSV for excel
write_excel_csv(x, path, na = "NA", append = FALSE, col_names = !append**)**

String to file
write_file(x, path, append = FALSE**)**

String vector to file, one element per line
write_lines(x,path, na = "NA", append = FALSE**)**

Object to RDS file
write_rds(x, path, compress = c("none", "gz", "bz2", "xz"), ...**)**

Tab delimited files
write_tsv(x, path, na = "NA", append = FALSE, col_names = !append**)**

Read Tabular Data - These functions share the common arguments:

read_*(file, col_names = TRUE, col_types = NULL, locale = default_locale(), na = c("", "NA"), quoted_na = TRUE, comment = "", trim_ws = TRUE, skip = 0, n_max = Inf, guess_max = min(1000, n_max), progress = interactive()**)**

```
a,b,c
1,2,3
4,5,NA
```

A	B	C
1	2	3
4	5	NA

Comma Delimited Files
read_csv("file.csv"**)**
To make file.csv run:
write_file(x = "a,b,c\n1,2,3\n4,5,NA", path = "file.csv")

```
a;b;c
1;2;3
4;5;NA
```

A	B	C
1	2	3
4	5	NA

Semi-colon Delimited Files
read_csv2("file2.csv"**)**
write_file(x = "a;b;c\n1;2;3\n4;5;NA", path = "file2.csv")

```
a|b|c
1|2|3
4|5|NA
```

A	B	C
1	2	3
4	5	NA

Files with Any Delimiter
read_delim("file.txt", delim = "|"**)**
write_file(x = "a|b|c\n1|2|3\n4|5|NA", path = "file.txt")

```
a b c
1 2 3
4 5 NA
```

A	B	C
1	2	3
4	5	NA

Fixed Width Files
read_fwf("file.fwf", col_positions = c(1, 3, 5)**)**
write_file(x = "a b c\n1 2 3\n4 5 NA", path = "file.fwf")

Tab Delimited Files
read_tsv("file.tsv"**)** Also **read_table().**
write_file(x = "a\tb\tc\n1\t2\t3\n4\t5\tNA", path = "file.tsv")

USEFUL ARGUMENTS

```
a,b,c
1,2,3
4,5,NA
```

Example file
write_file("a,b,c\n1,2,3\n4,5,NA","file.csv")
f <- "file.csv"

1	2	3
4	5	NA

Skip lines
read_csv(f, **skip = 1**)

A	B	C
1	2	3
4	5	NA

No header
read_csv(f, **col_names = FALSE**)

A	B	C
1	2	3

Read in a subset
read_csv(f, **n_max = 1**)

x	y	z
A	B	C
1	2	3
4	5	NA

Provide header
read_csv(f, **col_names = c("x", "y", "z")**)

A	B	C
NA	2	3
4	5	NA

Missing Values
read_csv(f, **na = c("1", ".")**)

Read Non-Tabular Data

Read a file into a single string
read_file(file, locale = default_locale()**)**

Read each line into its own string
read_lines(file, skip = 0, n_max = -1L, na = character(), locale = default_locale(), progress = interactive()**)**

Read Apache style log files
read_log(file, col_names = FALSE, col_types = NULL, skip = 0, n_max = -1, progress = interactive()**)**

Read a file into a raw vector
read_file_raw(file**)**

Read each line into a raw vector
read_lines_raw(file, skip = 0, n_max = -1L, progress = interactive()**)**

Data types

readr functions guess the types of each column and convert types when appropriate (but will NOT convert strings to factors automatically).

A message shows the type of each column in the result.

```
## Parsed with column specification:
## cols(
##   age = col_integer(),
##   sex = col_character(),
##   earn = col_double()
## )
```

age is an integer
sex is a character
earn is a double (numeric)

1. Use **problems()** to diagnose problems.
x <- read_csv("file.csv"); problems(x)

2. Use a col_ function to guide parsing.
- **col_guess()** - the default
- **col_character()**
- **col_double()**, **col_euro_double()**
- **col_datetime(**format = ""**)** Also **col_date(**format = ""**)**, **col_time(**format = ""**)**
- **col_factor(**levels, ordered = FALSE**)**
- **col_integer()**
- **col_logical()**
- **col_number()**, **col_numeric()**
- **col_skip()**

x <- read_csv("file.csv", col_types = cols(
A = col_double(),
B = col_logical(),
C = col_factor()))

3. Else, read in as character vectors then parse with a parse_ function.
- **parse_guess()**
- **parse_character()**
- **parse_datetime()** Also **parse_date()** and **parse_time()**
- **parse_double()**
- **parse_factor()**
- **parse_integer()**
- **parse_logical()**
- **parse_number()**

x$A <- parse_number(x$A)

RStudio® is a trademark of RStudio, Inc. • CC BY SA RStudio • info@rstudio.com • 844-448-1212 • rstudio.com • Learn more at tidyverse.org • readr 1.1.0 • tibble 1.2.12 • tidyr 0.6.0 • Updated: 2021–03

rstudio.com/resources/cheatsheets/

Ceci n'est pas une pipe.