R Stats Bootcamp

1.5 - Data frames

Megan Lewis

2024-11-28

R stats bootcamp - Module 1

Schedule:

- ~~Session 1: An introduction and script workflow~~
- ~~Session 2: R language~~
- ~~Session 3: R functions~~
- ~~Session 4: Data objects~~
- **Session 5: Data frames**
- Session 6: Data subsetting

Session 5 objectives:

- Common data file types
- Excel, data set up and the data dictionary
- Getting data into R
- Manipulating variables in the data frame
- Practice exercises

Data Frames in R

- A two dimensional data structure that organizes data into rows and column
- Can have different types of data inside
  - Though each column must be same data type

Common data file types

- csv: comma separated vales
    - Others available
- Excel
- Avoid (!) proprietary formats
- Data dictionary

Excel, data setup, and the data dictionary

💡 Like your room, data should be tidy

The first step in using R for data analysis is getting your data into R. The first step for getting your data into R is making your data tidy.

Tidy data concept

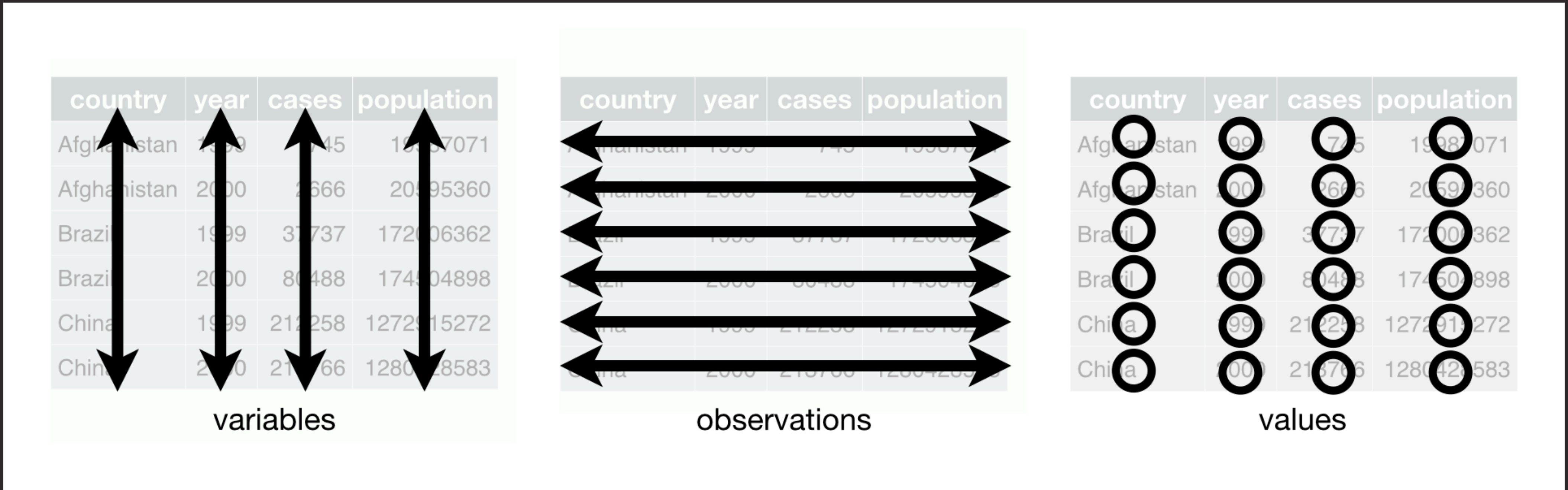
- Archives data into accessible format
- Makes data ‘transparent’ to others
- FAIR data principles
- “Tidy data” credit: Hadley Wickham
- Wickham, 2014

Tidy data concept

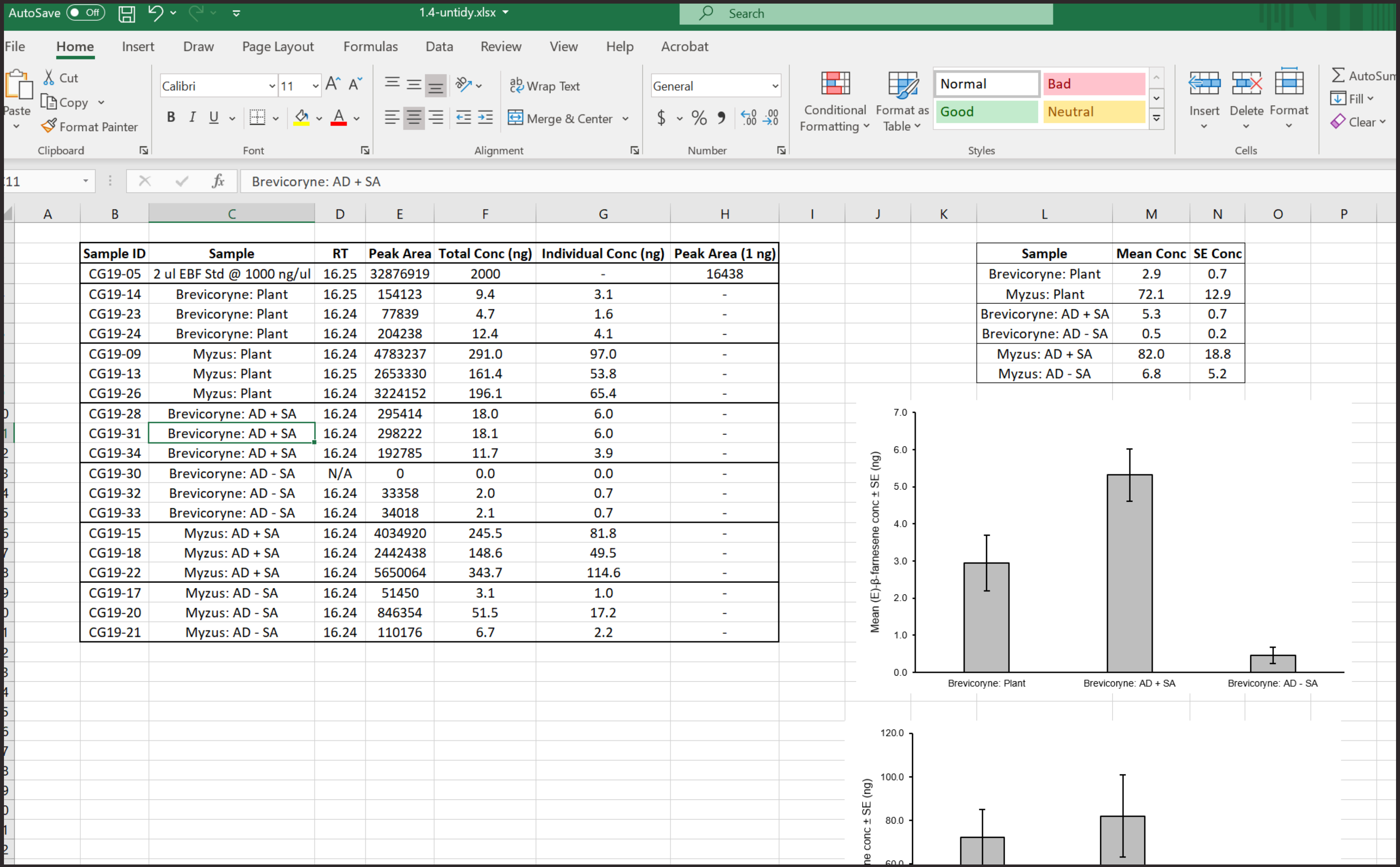
Essentials of tidy data:

- Each variable should be in a column
- Each independent observation should be in a row
- A data dictionary should be associated with the data set

Tidy data concept



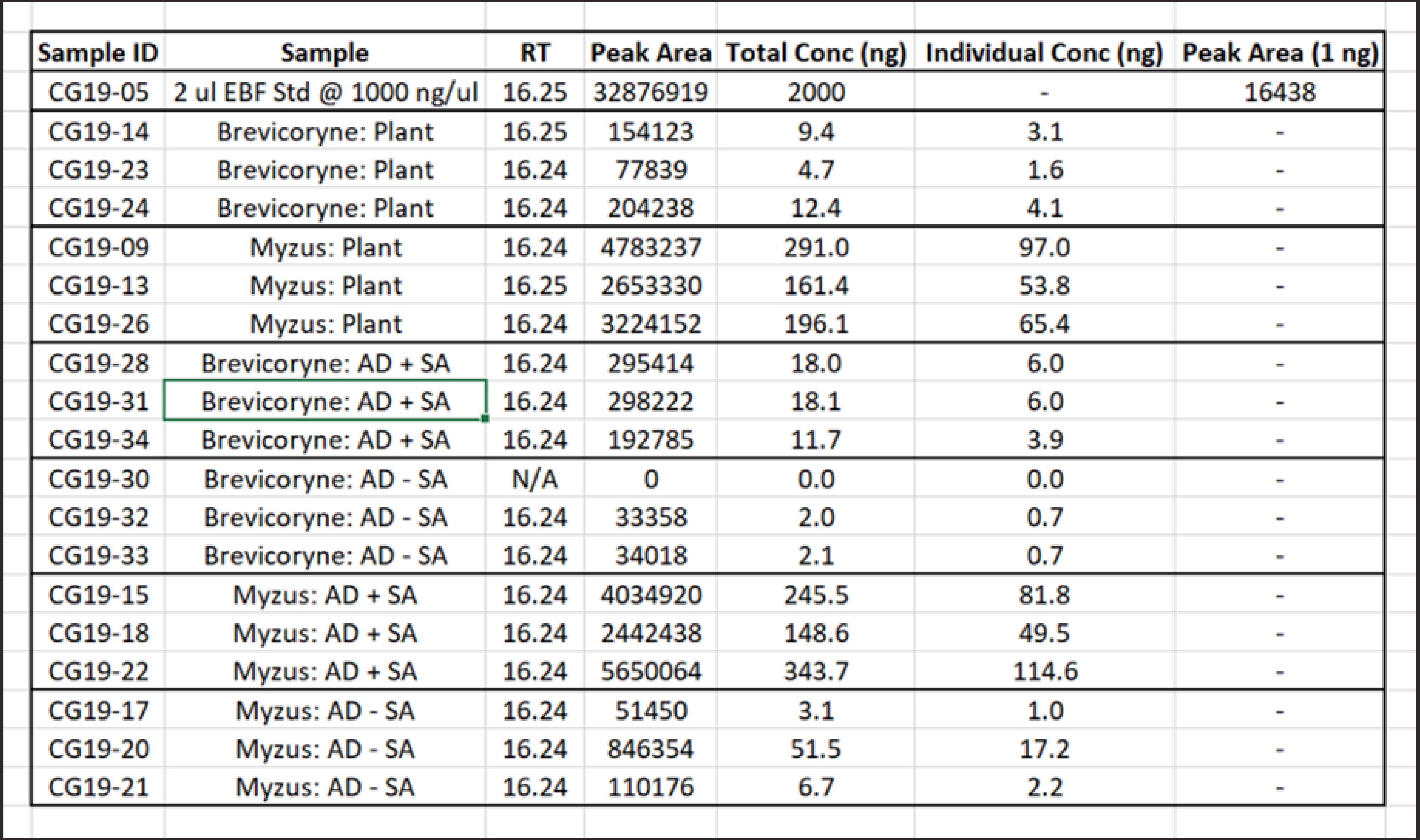
Untidy data



Sample ID	Sample	RT	Peak Area	Total Conc (ng)	Individual Conc (ng)	Peak Area (1 ng)
CG19-05	2 ul EBF Std @ 1000 ng/ul	16.25	32876919	2000	-	16438
CG19-14	Brevicoryne: Plant	16.25	154123	9.4	3.1	-
CG19-23	Brevicoryne: Plant	16.24	77839	4.7	1.6	-
CG19-24	Brevicoryne: Plant	16.24	204238	12.4	4.1	-
CG19-09	Myzus: Plant	16.24	4783237	291.0	97.0	-
CG19-13	Myzus: Plant	16.25	2653330	161.4	53.8	-
CG19-26	Myzus: Plant	16.24	3224152	196.1	65.4	-
CG19-28	Brevicoryne: AD + SA	16.24	295414	18.0	6.0	-
CG19-31	Brevicoryne: AD + SA	16.24	298222	18.1	6.0	-
CG19-34	Brevicoryne: AD + SA	16.24	192785	11.7	3.9	-
CG19-30	Brevicoryne: AD - SA	N/A	0	0.0	0.0	-
CG19-32	Brevicoryne: AD - SA	16.24	33358	2.0	0.7	-
CG19-33	Brevicoryne: AD - SA	16.24	34018	2.1	0.7	-
CG19-15	Myzus: AD + SA	16.24	4034920	245.5	81.8	-
CG19-18	Myzus: AD + SA	16.24	2442438	148.6	49.5	-
CG19-22	Myzus: AD + SA	16.24	5650064	343.7	114.6	-
CG19-17	Myzus: AD - SA	16.24	51450	3.1	1.0	-
CG19-20	Myzus: AD - SA	16.24	846354	51.5	17.2	-
CG19-21	Myzus: AD - SA	16.24	110176	6.7	2.2	-

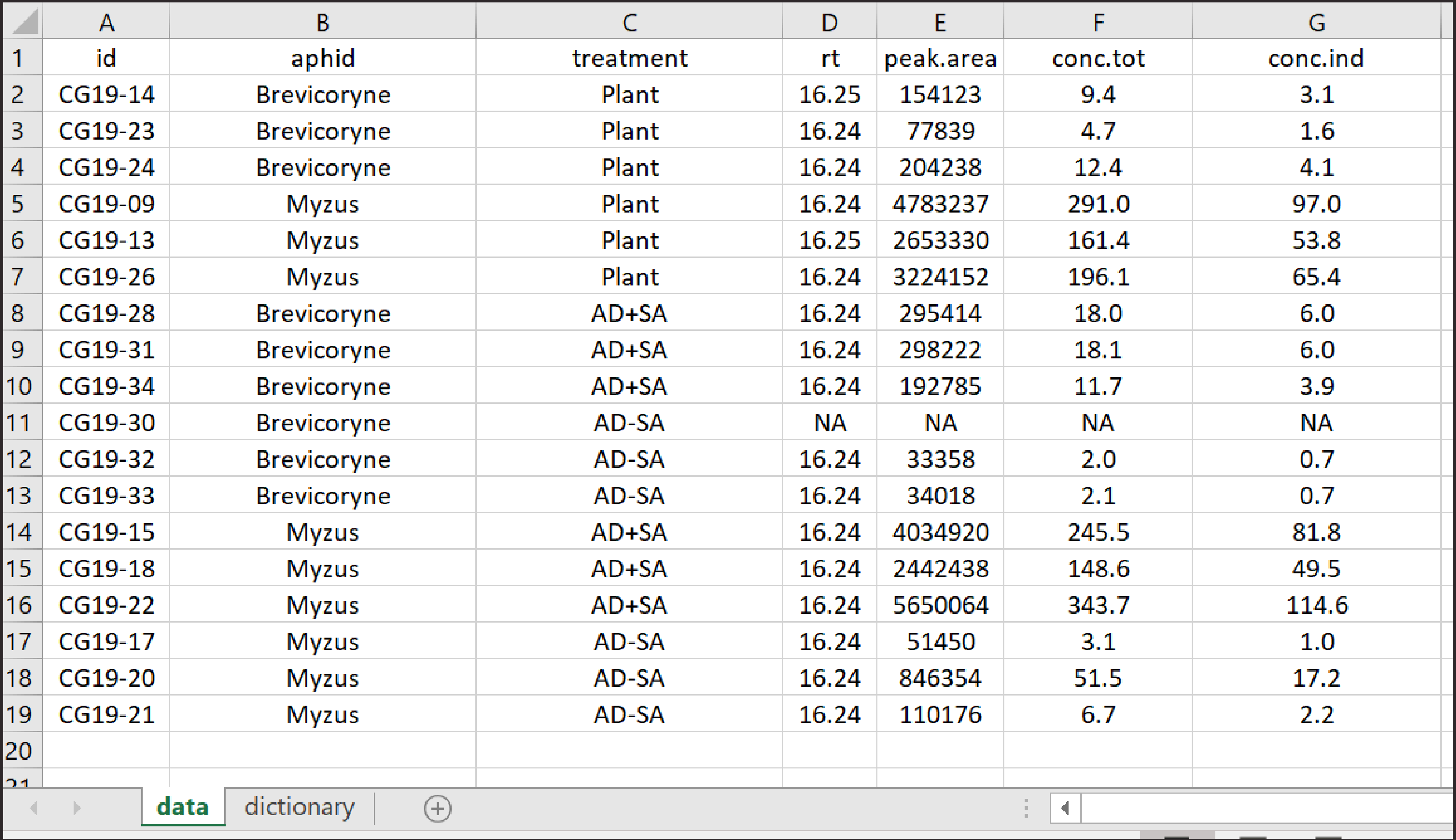
Sample	Mean Conc	SE Conc
Brevicoryne: Plant	2.9	0.7
Myzus: Plant	72.1	12.9
Brevicoryne: AD + SA	5.3	0.7
Brevicoryne: AD - SA	0.5	0.2
Myzus: AD + SA	82.0	18.8
Myzus: AD - SA	6.8	5.2

Untidy data



Sample ID	Sample	RT	Peak Area	Total Conc (ng)	Individual Conc (ng)	Peak Area (1 ng)
CG19-05	2 ul EBF Std @ 1000 ng/ul	16.25	32876919	2000	-	16438
CG19-14	Brevicoryne: Plant	16.25	154123	9.4	3.1	-
CG19-23	Brevicoryne: Plant	16.24	77839	4.7	1.6	-
CG19-24	Brevicoryne: Plant	16.24	204238	12.4	4.1	-
CG19-09	Myzus: Plant	16.24	4783237	291.0	97.0	-
CG19-13	Myzus: Plant	16.25	2653330	161.4	53.8	-
CG19-26	Myzus: Plant	16.24	3224152	196.1	65.4	-
CG19-28	Brevicoryne: AD + SA	16.24	295414	18.0	6.0	-
CG19-31	Brevicoryne: AD + SA	16.24	298222	18.1	6.0	-
CG19-34	Brevicoryne: AD + SA	16.24	192785	11.7	3.9	-
CG19-30	Brevicoryne: AD - SA	N/A	0	0.0	0.0	-
CG19-32	Brevicoryne: AD - SA	16.24	33358	2.0	0.7	-
CG19-33	Brevicoryne: AD - SA	16.24	34018	2.1	0.7	-
CG19-15	Myzus: AD + SA	16.24	4034920	245.5	81.8	-
CG19-18	Myzus: AD + SA	16.24	2442438	148.6	49.5	-
CG19-22	Myzus: AD + SA	16.24	5650064	343.7	114.6	-
CG19-17	Myzus: AD - SA	16.24	51450	3.1	1.0	-
CG19-20	Myzus: AD - SA	16.24	846354	51.5	17.2	-
CG19-21	Myzus: AD - SA	16.24	110176	6.7	2.2	-

Tidy data



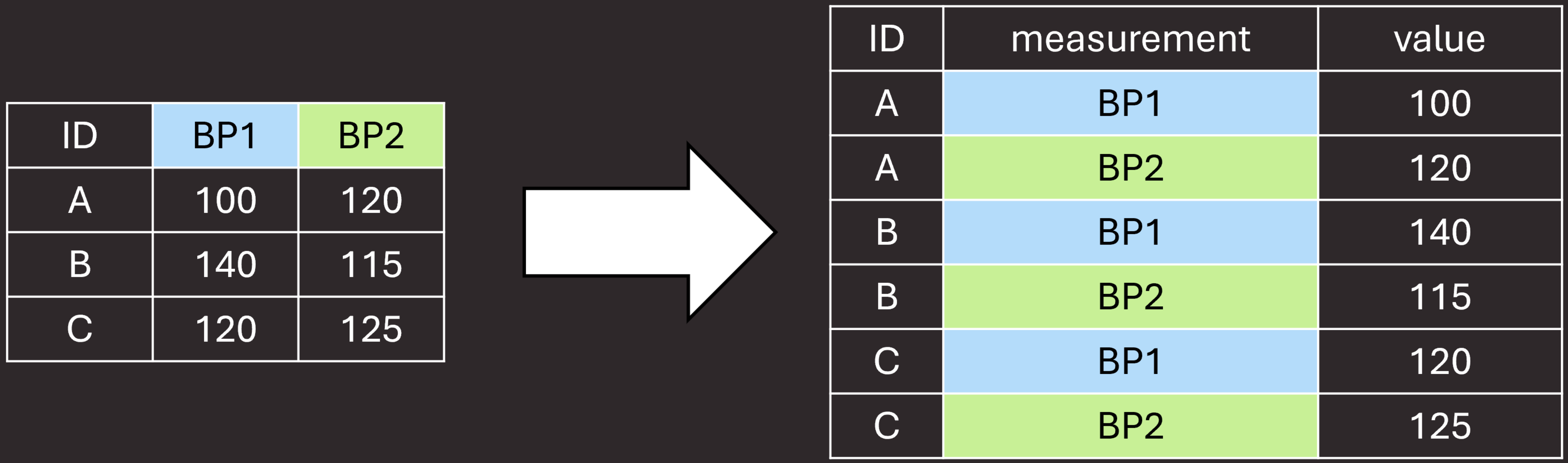
	A	B	C	D	E	F	G
1	id	aphid	treatment	rt	peak.area	conc.tot	conc.ind
2	CG19-14	Brevicoryne	Plant	16.25	154123	9.4	3.1
3	CG19-23	Brevicoryne	Plant	16.24	77839	4.7	1.6
4	CG19-24	Brevicoryne	Plant	16.24	204238	12.4	4.1
5	CG19-09	Myzus	Plant	16.24	4783237	291.0	97.0
6	CG19-13	Myzus	Plant	16.25	2653330	161.4	53.8
7	CG19-26	Myzus	Plant	16.24	3224152	196.1	65.4
8	CG19-28	Brevicoryne	AD+SA	16.24	295414	18.0	6.0
9	CG19-31	Brevicoryne	AD+SA	16.24	298222	18.1	6.0
10	CG19-34	Brevicoryne	AD+SA	16.24	192785	11.7	3.9
11	CG19-30	Brevicoryne	AD-SA	NA	NA	NA	NA
12	CG19-32	Brevicoryne	AD-SA	16.24	33358	2.0	0.7
13	CG19-33	Brevicoryne	AD-SA	16.24	34018	2.1	0.7
14	CG19-15	Myzus	AD+SA	16.24	4034920	245.5	81.8
15	CG19-18	Myzus	AD+SA	16.24	2442438	148.6	49.5
16	CG19-22	Myzus	AD+SA	16.24	5650064	343.7	114.6
17	CG19-17	Myzus	AD-SA	16.24	51450	3.1	1.0
18	CG19-20	Myzus	AD-SA	16.24	846354	51.5	17.2
19	CG19-21	Myzus	AD-SA	16.24	110176	6.7	2.2
20							

Tidy data concept

ID	BP1	BP2
A	100	120
B	140	115
C	120	125

ID	measurement	value
A	BP1	100
A	BP2	120
B	BP1	140
B	BP2	115
C	BP1	120
C	BP2	125

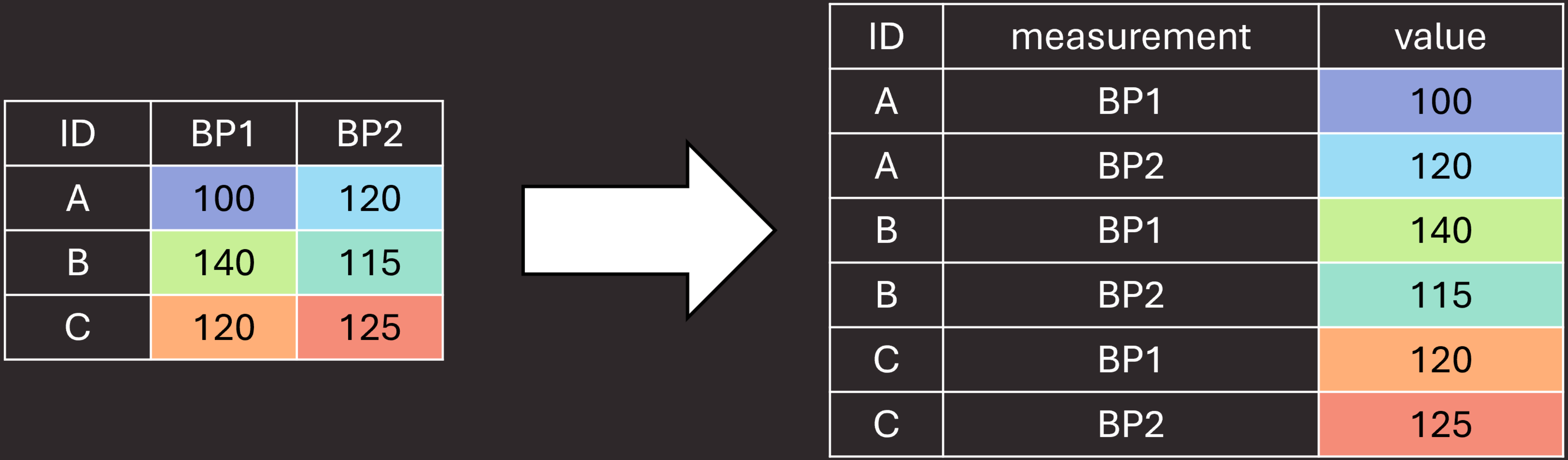
Tidy data concept



ID	BP1	BP2
A	100	120
B	140	115
C	120	125

ID	measurement	value
A	BP1	100
A	BP2	120
B	BP1	140
B	BP2	115
C	BP1	120
C	BP2	125

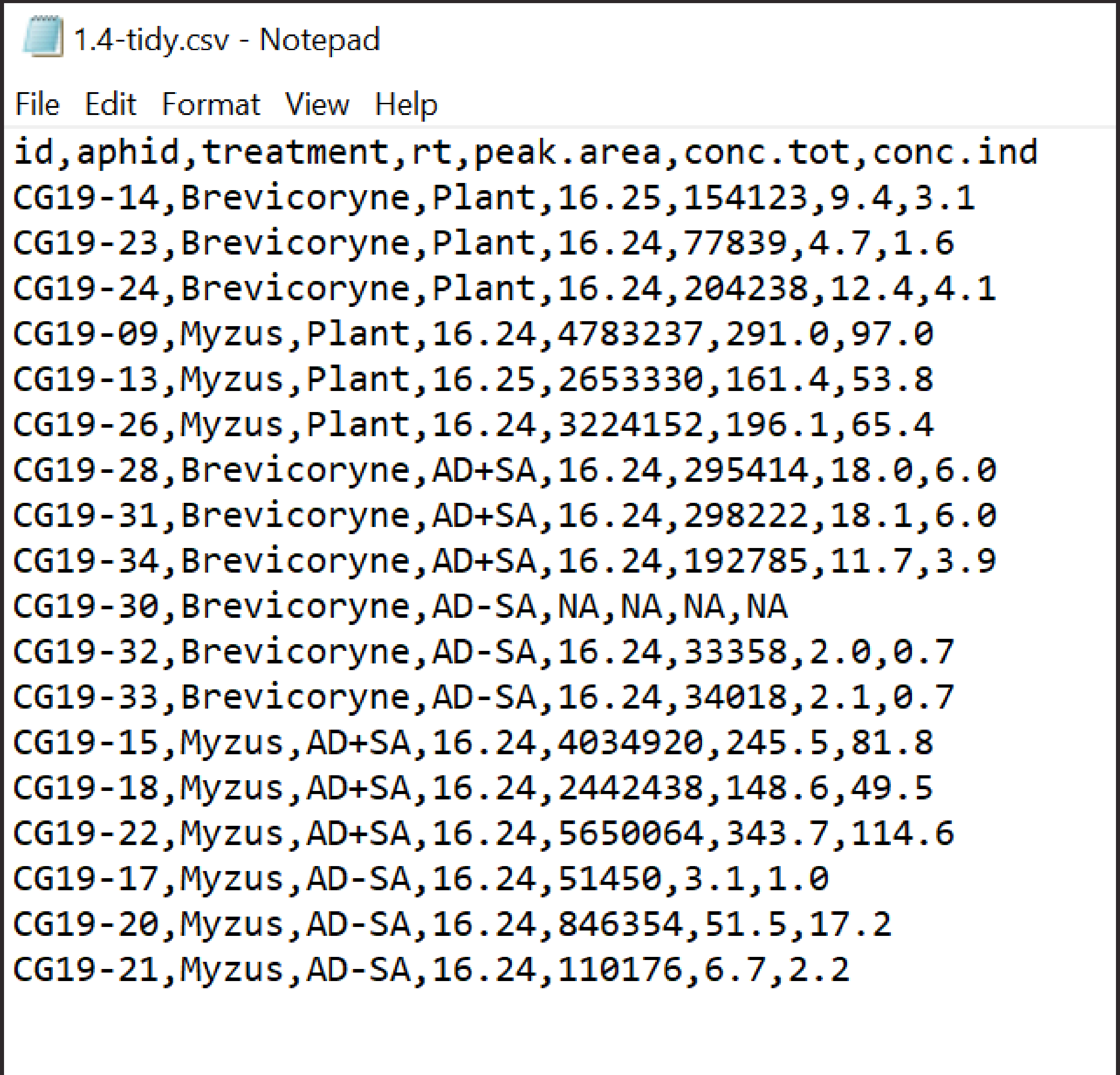
Tidy data concept



ID	BP1	BP2
A	100	120
B	140	115
C	120	125

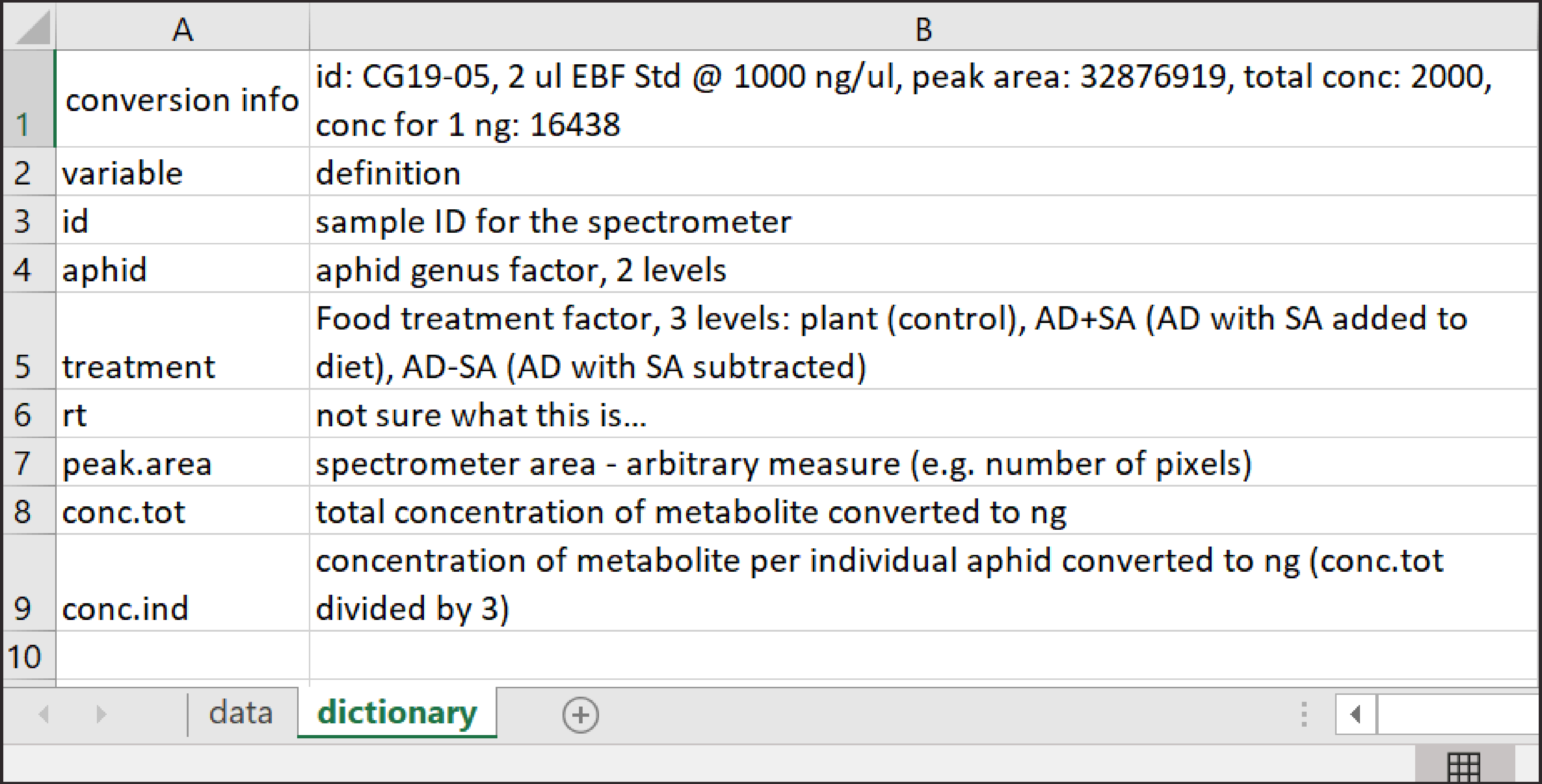
ID	measurement	value
A	BP1	100
A	BP2	120
B	BP1	140
B	BP2	115
C	BP1	120
C	BP2	125

Tidy csv



```
id,aphid,treatment,rt,peak.area,conc.tot,conc.ind
CG19-14,Brevicoryne,Plant,16.25,154123,9.4,3.1
CG19-23,Brevicoryne,Plant,16.24,77839,4.7,1.6
CG19-24,Brevicoryne,Plant,16.24,204238,12.4,4.1
CG19-09,Myzus,Plant,16.24,4783237,291.0,97.0
CG19-13,Myzus,Plant,16.25,2653330,161.4,53.8
CG19-26,Myzus,Plant,16.24,3224152,196.1,65.4
CG19-28,Brevicoryne,AD+SA,16.24,295414,18.0,6.0
CG19-31,Brevicoryne,AD+SA,16.24,298222,18.1,6.0
CG19-34,Brevicoryne,AD+SA,16.24,192785,11.7,3.9
CG19-30,Brevicoryne,AD-SA,NA,NA,NA,NA
CG19-32,Brevicoryne,AD-SA,16.24,33358,2.0,0.7
CG19-33,Brevicoryne,AD-SA,16.24,34018,2.1,0.7
CG19-15,Myzus,AD+SA,16.24,4034920,245.5,81.8
CG19-18,Myzus,AD+SA,16.24,2442438,148.6,49.5
CG19-22,Myzus,AD+SA,16.24,5650064,343.7,114.6
CG19-17,Myzus,AD-SA,16.24,51450,3.1,1.0
CG19-20,Myzus,AD-SA,16.24,846354,51.5,17.2
CG19-21,Myzus,AD-SA,16.24,110176,6.7,2.2
```

Data dictionary



	A	B
1	conversion info	id: CG19-05, 2 ul EBF Std @ 1000 ng/ul, peak area: 32876919, total conc: 2000, conc for 1 ng: 16438
2	variable	definition
3	id	sample ID for the spectrometer
4	aphid	aphid genus factor, 2 levels
5	treatment	Food treatment factor, 3 levels: plant (control), AD+SA (AD with SA added to diet), AD-SA (AD with SA subtracted)
6	rt	not sure what this is...
7	peak.area	spectrometer area - arbitrary measure (e.g. number of pixels)
8	conc.tot	total concentration of metabolite converted to ng
9	conc.ind	concentration of metabolite per individual aphid converted to ng (conc.tot divided by 3)
10		

Getting data into R

Choices

- File > Import data
- read.csv()
- read.table()
- readxl::read_excel()
- openxlsx::read.xlsx()

Getting data into R - Working directory

Working directories

Best practice when working with files is to formally set your “working directory”. Basically, this tells R where your input (i.e. data) and output (like scripts or figures) files should be.

Working directories and Windows

R file paths use the forward slash symbol “/” to separate file names. A very important step for Windows users when setting the working directory in R is to change the Windows default “” for forward slashes…

Demo

Manipulating variables in the Data frame

R space

Now that there is a data frame in your working environment, we can start working with the variables. This is a good time to think about the "R Space" metaphor. You are floating in `R` Space and you can see a data frame called `my_data`. You cannot see inside the container, so we will look at methods of accessing the data inside by name…

Manipulating variables in the Data frame

- `class()`
- `names()`
- `str()`
- `attach()`
- indexes `[ , ]`

Demo

Practice exercises

Practice exercise 01

- Download the butterfly data file and place it in a working directory.
- Set your working directory.
- Read the data file and place it in a data frame object named `data1`
- After examining the data, use `mean()` to calculate the mean of the variable `length` and report the results in a comment to 2dp accuracy.

Practice exercise 02

- Show the code to convert the `diet` variable to an ordinal factor with the order “control” > “enhanced” and the `sex` variable to a plain categorical factor.

Practice exercise 03

Show code for two different variations of using only the `[ , ]` operator with your data frame to show the following:

	diet	length
8	control	6
9	control	7
10	control	6
11	enhanced	8
12	enhanced	7
13	enhanced	9

Practice exercise 04

- Show code to read in a comma separated values data file that does not have a header (i.e., does not have a first row containing variable names)

Practice exercise 05

- Describe in your own words what the `attach()` function does

Practice exercise 06

- Write a plausible practice question involving any aspect of manipulation of a data frame.