UC San Diego

DISCUSSION SECTIONS (A01 & A02)
October 19, 2022

Ashwin Mishra, TA
Jiesen Zhang, IA   Lindsey Gu, IA

Tidy Data

1. Each variable you measure should be in a single column

ID	LastName	Age	Occupation
1842	Smith	55	Doctor
2432	Doe	34	Data Scientist
3417	Jordan	19	Student

Tidy Data

2.   Every observation of a variable should be in a different row

ID	LastName	Age	Occupation
1842	Smith	55	Doctor
2432	Doe	34	Data Scientist
3417	Jordan	19	Student

Tidy Data

3. There should be one table for each type of data

Demographic data

ID	LastName	Age	Occupation
1842	Smith	55	Doctor
2432	Doe	34	Data Scientist
3417	Jordan	19	Student

Medical data

ID	Height_inch	Weight_lbs	Insulin
1842	65	180	0.60
2432	75	215	1.46
3417	62	124	0.72

Tidy Data

4.   If you have multiple tables, they should include a column in each with the same column label that allows them to be joined or merged.

Demographic data

ID	LastName	Age	Occupation
1842	Smith	55	Doctor
2432	Doe	34	Data Scientist
3417	Jordan	19	Student

Medical data

ID	Height_inch	Weight_lbs	Insulin
1842	65	180	0.60
2432	75	215	1.46
3417	62	124	0.72

Rules for tidy data

1. Be consistent

ID	LastName	Sex	City
37183	Smith	Male	San Diego
43842	Jordan	Female	Los Angeles

ID_user	Sex	Height_inch
37183	M	65
43842	F	75

Rules for tidy data

2.   Proper naming conventions

	Do	Don't
Avoid Extra Spaces	"male"	"male "
Use underscores not spaces	doctor_visit_1	doctor visit 1
Choose meaningful names	doctor_visit	"Var1"

Rules for tidy data
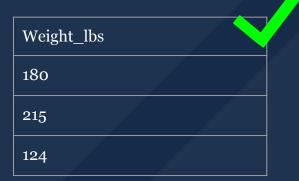
3.   Write dates as YYYY-MM-DD (ISO 8601)

	Do	Don't
Use ISO-8601 standard	2018-02-27	2/27 or 2_27_2018 or Feb 27

Rules for tidy data

4. No empty cells

ID	LastName	Sex	
37183	Smith	Male	San Diego
43842	Jordan		Los Angeles

Rules for tidy data

5. Put only one thing in a cell

Weight_lbs
180
215
124

✅

Weight
180 lbs
215 lbs
124 lbs

❌

Rules for tidy data

6.   Don't use font color/highlighting as data

glucose
149.3
95.3
1.1
108.5

glucose	outlier
149.3	False
95.3	False
1.1	True
108.5	False