Andrew Gard - equitable.equations@gmail.com

Tidy Data

There are many possible ways to record data in a spreadsheet.

There are many possible ways to record data in a spreadsheet. Different formats may be appropriate for different applications, but there is one that is particularly convenient in data science: **tidy data**.

There are many possible ways to record data in a spreadsheet. Different formats may be appropriate for different applications, but there is one that is particularly convenient in data science: **tidy data**. Tidy data obeys three simple principles.

There are many possible ways to record data in a spreadsheet. Different formats may be appropriate for different applications, but there is one that is particularly convenient in data science: **tidy data**. Tidy data obeys three simple principles.

1. One observation per row

There are many possible ways to record data in a spreadsheet. Different formats may be appropriate for different applications, but there is one that is particularly convenient in data science: **tidy data**. Tidy data obeys three simple principles.

1. One observation per row
2. One variable per column

There are many possible ways to record data in a spreadsheet. Different formats may be appropriate for different applications, but there is one that is particularly convenient in data science: **tidy data**. Tidy data obeys three simple principles.

1. One observation per row
2. One variable per column
3. One type of observation per data set

There are many possible ways to record data in a spreadsheet. Different formats may be appropriate for different applications, but there is one that is particularly convenient in data science: **tidy data**. Tidy data obeys three simple principles.

1. One observation per row
2. One variable per column
3. One type of observation per data set

In this context, a *variable* is something that is measured across all experimental units and an *observation* is the collection of all variable measurements for a single unit.

There are many possible ways to record data in a spreadsheet. Different formats may be appropriate for different applications, but there is one that is particularly convenient in data science: **tidy data**. Tidy data obeys three simple principles.

1. One observation per row
2. One variable per column
3. One type of observation per data set

In this context, a *variable* is something that is measured across all experimental units and an *observation* is the collection of all variable measurements for a single unit.

An immediate advantage to tidy data is that new observations (for instance, subjects in a medical trial) can easily be added as rows at the bottom of an existing data set, and new variables (like subject ID numbers) can easily be added as columns at the right.

The `mtcars` data set, which comes pre-loaded with *R*, is tidy.

mtcars

Filter

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4

Showing 1 to 10 of 32 entries, 11 total columns

Each row represents a single car. Each column represents a single characteristic of those cars.

Ideally, every tidy data set should be accompanied by a **data dictionary** that explains the meaning of each variable (column) in the data set, including the units of measure.

Ideally, every tidy data set should be accompanied by a **data dictionary** that explains the meaning of each variable (column) in the data set, including the units of measure. A good data dictionary should also include information about the set itself: when, how, and by whom the data was recorded.

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 (numeric) variables.

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	Engine (0 = V-shaped, 1 = straight)
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

The `diamonds` data set in R's `ggplot2` package is also tidy.

diamonds

Filter

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

Showing 1 to 10 of 53,940 entries, 10 total columns

Each row represents a single round-cut diamond. Each column represents a single characteristic of those diamonds.

On the other hand, the `construction` data set in Rś `tidyr` package is not tidy.

construction

Filter

	Year	Month	1 unit	2 to 4 units	5 units or more	Northeast	Midwest	South	West
1	2018	January	859	NA	348	114	169	596	339
2	2018	February	882	NA	400	138	160	655	336
3	2018	March	862	NA	356	150	154	595	330
4	2018	April	797	NA	447	144	196	613	304
5	2018	May	875	NA	364	90	169	673	319
6	2018	June	867	NA	342	76	170	610	360
7	2018	July	829	NA	360	108	183	594	310
8	2018	August	939	NA	286	90	205	649	286
9	2018	September	835	NA	304	117	175	560	296

Showing 1 to 9 of 9 entries, 9 total columns

On the other hand, the `construction` data set in Rś `tidyr` package is not tidy.

construction

Filter

	Year	Month	1 unit	2 to 4 units	5 units or more	Northeast	Midwest	South	West
1	2018	January	859	NA	348	114	169	596	339
2	2018	February	882	NA	400	138	160	655	336
3	2018	March	862	NA	356	150	154	595	330
4	2018	April	797	NA	447	144	196	613	304
5	2018	May	875	NA	364	90	169	673	319
6	2018	June	867	NA	342	76	170	610	360
7	2018	July	829	NA	360	108	183	594	310
8	2018	August	939	NA	286	90	205	649	286
9	2018	September	835	NA	304	117	175	560	296

Showing 1 to 9 of 9 entries, 9 total columns

In this data set, two variables, `Number of units` and `Region` have been recorded across multiple columns.

Untidy data isn't necessarily bad. Real-world spreadsheets often use field-specific conventions to aid with data entry, readability, or both. These conventions might include annotations, special formatting, or multiple types of observations in a single set, to name just a few.

Untidy data isn't necessarily bad. Real-world spreadsheets often use field-specific conventions to aid with data entry, readability, or both. These conventions might include annotations, special formatting, or multiple types of observations in a single set, to name just a few.

However, the tidy format is perfectly suited to data science, which seeks to explore relationships between variables across large collections of observations. Whether you're creating data visualization or building sophisticated predictive models, having variables in individual columns and observations in individual rows will make your life just a little bit easier.

A particularly common form of untidy data is the **contingency table**, which shows counts for various combinations of categorical variables.

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

(table taken from *OpenIntro Statistics* by Diez, Çetinkaya-Rundel, and Barr)

A particularly common form of untidy data is the **contingency table**, which shows counts for various combinations of categorical variables.

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

(table taken from *OpenIntro Statistics* by Diez, Çetinkaya-Rundel, and Barr)

This table includes variables in both its rows and columns. Moreover, its cells refer to a third piece of information entirely, the number of loan applicants with each combination of variable values.

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

A more tidy version of this table would have three columns: `homeownership` `app_type`, and `count`.

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

A more tidy version of this table would have three columns: `homeownership` `app_type`, and `count`.

app_type	homeownership	count
individual	rent	3496
individual	mortgage	3839
individual	own	1170
joint	rent	362
joint	mortgage	950
joint	own	183

		homeownership			
		rent	mortgage	own	Total
app_type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

A more tidy version of this table would have three columns: `homeownership` `app_type`, and `count`.

app_type	homeownership	count
individual	rent	3496
individual	mortgage	3839
individual	own	1170
joint	rent	362
joint	mortgage	950
joint	own	183

A fully tidy version of this would include 10,000 rows, one for each loan application in the data set, and two columns, `homeownership` and `app_type`.