Tidy Data

Tidy Data

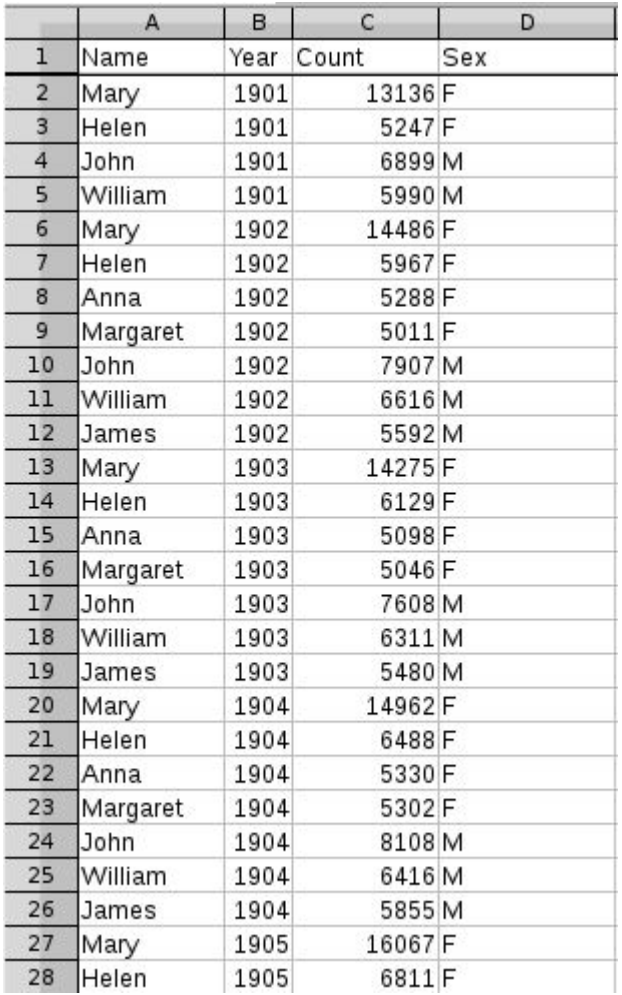
Tidy data satisfy three requirements:

1. Each variable forms a column
2. Each observation forms a row
3. A set of observations with values for most, if not all, variables forms a table

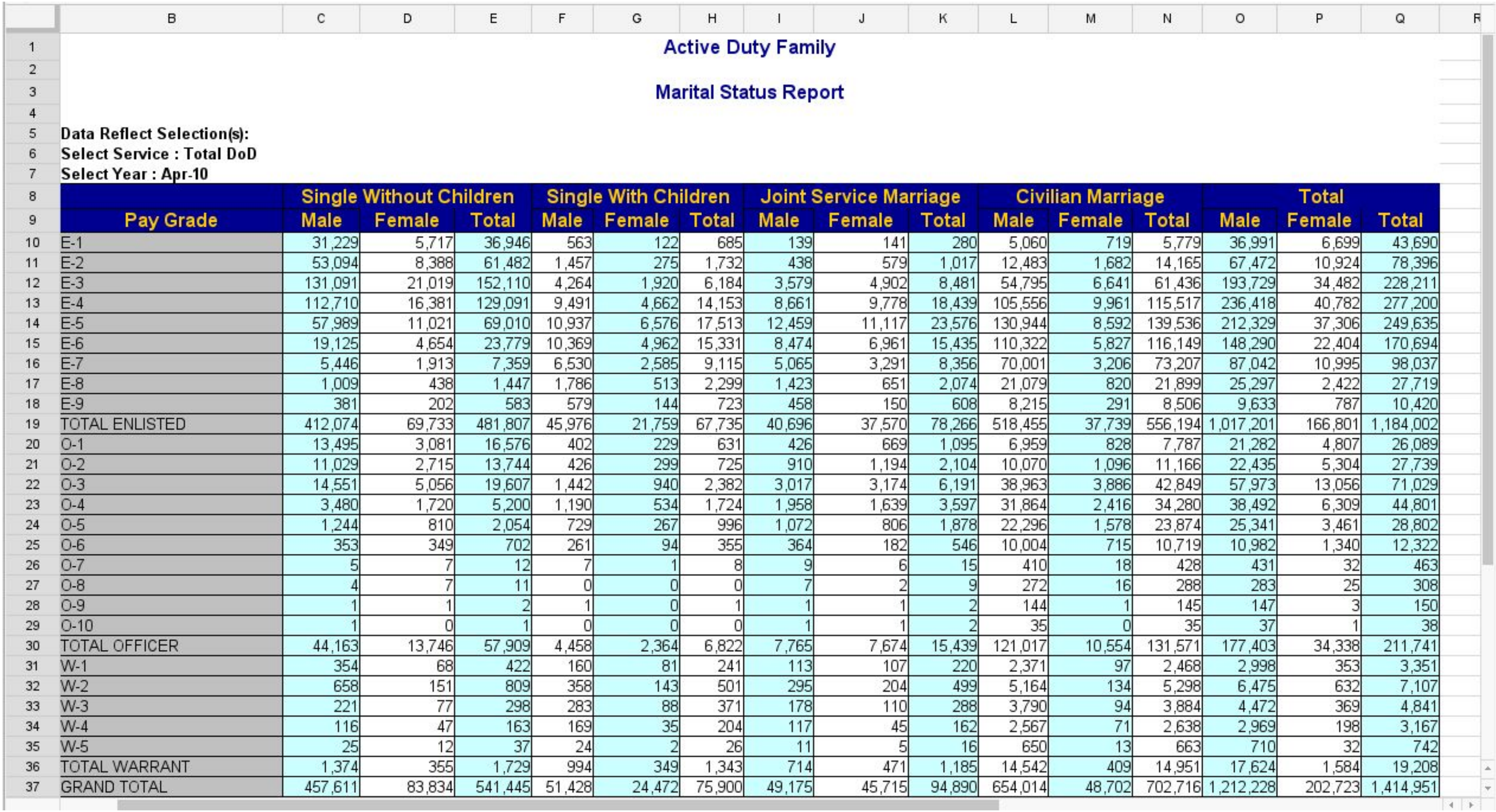
When data are arranged this way, we can easily summarize observations: e.g., total some variable

Example

- Column headings are variables:
  - Name, Year, Count, Sex
- Rows are observations:
  - Values for Name (Mary, Helen, etc.)
  - Values for Year (1901, 1902, etc.)
  - Values for Count (natural numbers)
  - Values for Sex (F, M)



	A	B	C	D
1	Name	Year	Count	Sex
2	Mary	1901	13136	F
3	Helen	1901	5247	F
4	John	1901	6899	M
5	William	1901	5990	M
6	Mary	1902	14486	F
7	Helen	1902	5967	F
8	Anna	1902	5288	F
9	Margaret	1902	5011	F
10	John	1902	7907	M
11	William	1902	6616	M
12	James	1902	5592	M
13	Mary	1903	14275	F
14	Helen	1903	6129	F
15	Anna	1903	5098	F
16	Margaret	1903	5046	F
17	John	1903	7608	M
18	William	1903	6311	M
19	James	1903	5480	M
20	Mary	1904	14962	F
21	Helen	1904	6488	F
22	Anna	1904	5330	F
23	Margaret	1904	5302	F
24	John	1904	8108	M
25	William	1904	6416	M
26	James	1904	5855	M
27	Mary	1905	16067	F
28	Helen	1905	6811	F



Active Duty Family

Marital Status Report

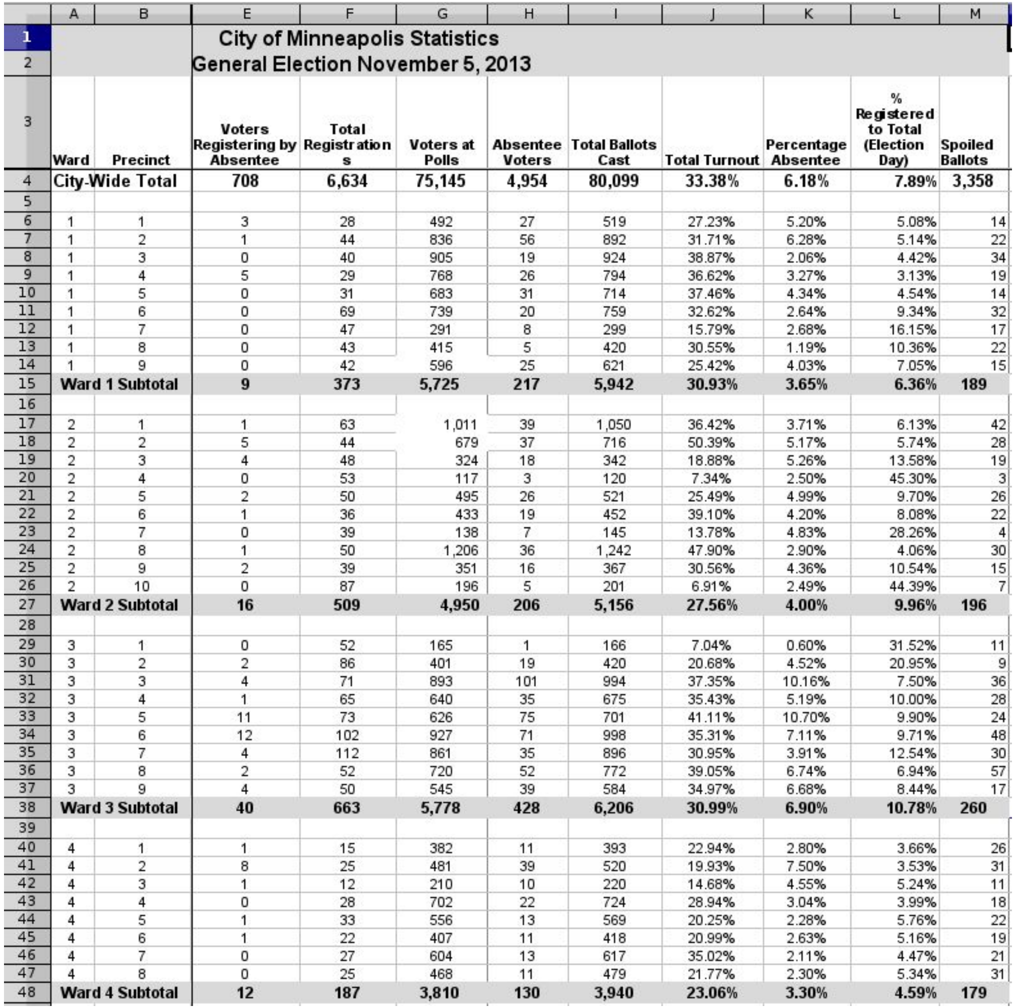
Data Reflect Selection(s):
Select Service : Total DoD
Select Year : Apr-10

	Single Without Children			Single With Children			Joint Service Marriage			Civilian Marriage			Total		
Pay Grade	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
E-1	31,229	5,717	36,946	563	122	685	139	141	280	5,060	719	5,779	36,991	6,699	43,690
E-2	53,094	8,388	61,482	1,457	275	1,732	438	579	1,017	12,483	1,682	14,165	67,472	10,924	78,396
E-3	131,091	21,019	152,110	4,264	1,920	6,184	3,579	4,902	8,481	54,795	6,641	61,436	193,729	34,482	228,211
E-4	112,710	16,381	129,091	9,491	4,662	14,153	8,661	9,778	18,439	105,556	9,961	115,517	236,418	40,782	277,200
E-5	57,989	11,021	69,010	10,937	6,576	17,513	12,459	11,117	23,576	130,944	8,592	139,536	212,329	37,306	249,635
E-6	19,125	4,654	23,779	10,369	4,962	15,331	8,474	6,961	15,435	110,322	5,827	116,149	148,290	22,404	170,694
E-7	5,446	1,913	7,359	6,530	2,585	9,115	5,065	3,291	8,356	70,001	3,206	73,207	87,042	10,995	98,037
E-8	1,009	438	1,447	1,786	513	2,299	1,423	651	2,074	21,079	820	21,899	25,297	2,422	27,719
E-9	381	202	583	579	144	723	458	150	608	8,215	291	8,506	9,633	787	10,420
TOTAL ENLISTED	412,074	69,733	481,807	45,976	21,759	67,735	40,696	37,570	78,266	518,455	37,739	556,194	1,017,201	166,801	1,184,002
O-1	13,495	3,081	16,576	402	229	631	426	669	1,095	6,959	828	7,787	21,282	4,807	26,089
O-2	11,029	2,715	13,744	426	299	725	910	1,194	2,104	10,070	1,096	11,166	22,435	5,304	27,739
O-3	14,551	5,056	19,607	1,442	940	2,382	3,017	3,174	6,191	38,963	3,886	42,849	57,973	13,056	71,029
O-4	3,480	1,720	5,200	1,190	534	1,724	1,958	1,639	3,597	31,864	2,416	34,280	38,492	6,309	44,801
O-5	1,244	810	2,054	729	267	996	1,072	806	1,878	22,296	1,578	23,874	25,341	3,461	28,802
O-6	353	349	702	261	94	355	364	182	546	10,004	715	10,719	10,982	1,340	12,322
O-7	5	7	12	7	1	8	9	6	15	410	18	428	431	32	463
O-8	4	7	11	0	0	0	7	2	9	272	16	288	283	25	308
O-9	1	1	2	1	0	1	1	1	2	144	1	145	147	3	150
O-10	1	0	1	0	0	0	1	1	2	35	0	35	37	1	38
TOTAL OFFICER	44,163	13,746	57,909	4,458	2,364	6,822	7,765	7,674	15,439	121,017	10,554	131,571	177,403	34,338	211,741
W-1	354	68	422	160	81	241	113	107	220	2,371	97	2,468	2,998	353	3,351
W-2	658	151	809	358	143	501	295	204	499	5,164	134	5,298	6,475	632	7,107
W-3	221	77	298	283	88	371	178	110	288	3,790	94	3,884	4,472	369	4,841
W-4	116	47	163	169	35	204	117	45	162	2,567	71	2,638	2,969	198	3,167
W-5	25	12	37	24	2	26	11	5	16	650	13	663	710	32	742
TOTAL WARRANT	1,374	355	1,729	994	349	1,343	714	471	1,185	14,542	409	14,951	17,624	1,584	19,208
GRAND TOTAL	457,611	83,834	541,445	51,428	24,472	75,900	49,175	45,715	94,890	654,014	48,702	702,716	1,212,228	202,723	1,414,951

Data source

Why aren’t these data tidy?

- Which column headings are values?
- Which row headings are values?
- Which columns are a summary of other columns?
- Which rows are a summary of other rows?
- What should the variables be?
- What should the observations be?

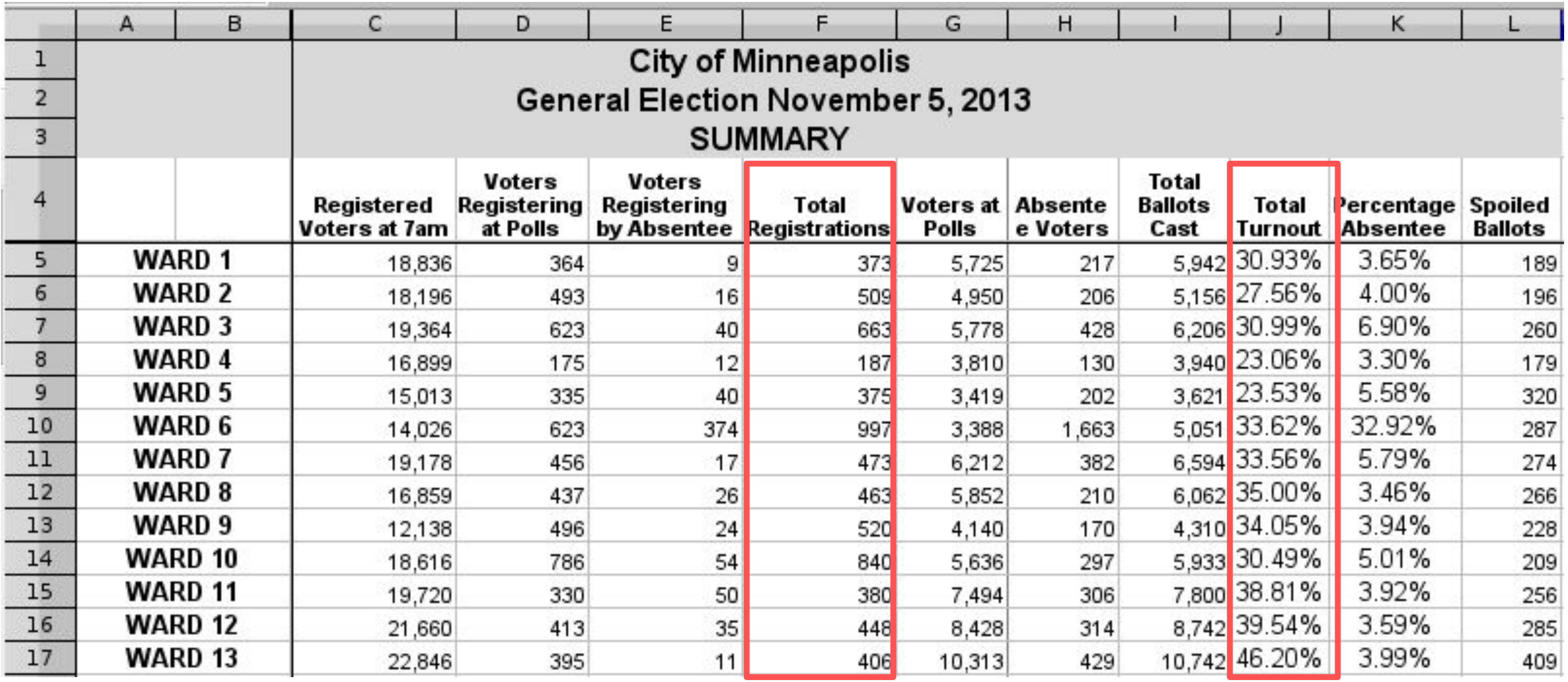


	A	B	E	F	G	H	I	J	K	L	M
1			City of Minneapolis Statistics								
2			General Election November 5, 2013								
3	Ward	Precinct	Voters Registering by Absentee	Total Registrations	Voters at Polls	Absentee Voters	Total Ballots Cast	Total Turnout	Percentage Absentee	% Registered to Total (Election Day)	Spoiled Ballots
4	City-Wide Total		708	6,634	75,145	4,954	80,099	33.38%	6.18%	7.89%	3,358
5											
6	1	1	3	28	492	27	519	27.23%	5.20%	5.08%	14
7	1	2	1	44	836	56	892	31.71%	6.28%	5.14%	22
8	1	3	0	40	905	19	924	38.87%	2.06%	4.42%	34
9	1	4	5	29	768	26	794	36.62%	3.27%	3.13%	19
10	1	5	0	31	683	31	714	37.46%	4.34%	4.54%	14
11	1	6	0	69	739	20	759	32.62%	2.64%	9.34%	32
12	1	7	0	47	291	8	299	15.79%	2.68%	16.15%	17
13	1	8	0	43	415	5	420	30.55%	1.19%	10.36%	22
14	1	9	0	42	596	25	621	25.42%	4.03%	7.05%	15
15	**Ward 1 Subtotal**		**9**	**373**	**5,725**	**217**	**5,942**	**30.93%**	**3.65%**	**6.36%**	**189**
16											
17	2	1	1	63	1,011	39	1,050	36.42%	3.71%	6.13%	42
18	2	2	5	44	679	37	716	50.39%	5.17%	5.74%	28
19	2	3	4	48	324	18	342	18.88%	5.26%	13.58%	19
20	2	4	0	53	117	3	120	7.34%	2.50%	45.30%	3
21	2	5	2	50	495	26	521	25.49%	4.99%	9.70%	26
22	2	6	1	36	433	19	452	39.10%	4.20%	8.08%	22
23	2	7	0	39	138	7	145	13.78%	4.83%	28.26%	4
24	2	8	1	50	1,206	36	1,242	47.90%	2.90%	4.06%	30
25	2	9	2	39	351	16	367	30.56%	4.36%	10.54%	15
26	2	10	0	87	196	5	201	6.91%	2.49%	44.39%	7
27	**Ward 2 Subtotal**		**16**	**509**	**4,950**	**206**	**5,156**	**27.56%**	**4.00%**	**9.96%**	**196**
28											
29	3	1	0	52	165	1	166	7.04%	0.60%	31.52%	11
30	3	2	2	86	401	19	420	20.68%	4.52%	20.95%	9
31	3	3	4	71	893	101	994	37.35%	10.16%	7.50%	36
32	3	4	1	65	640	35	675	35.43%	5.19%	10.00%	28
33	3	5	11	73	626	75	701	41.11%	10.70%	9.90%	24
34	3	6	12	102	927	71	998	35.31%	7.11%	9.71%	48
35	3	7	4	112	861	35	896	30.95%	3.91%	12.54%	30
36	3	8	2	52	720	52	772	39.05%	6.74%	6.94%	57
37	3	9	4	50	545	39	584	34.97%	6.68%	8.44%	17
38	**Ward 3 Subtotal**		**40**	**663**	**5,778**	**428**	**6,206**	**30.99%**	**6.90%**	**10.78%**	**260**
39											
40	4	1	1	15	382	11	393	22.94%	2.80%	3.66%	26
41	4	2	8	25	481	39	520	19.93%	7.50%	3.53%	31
42	4	3	1	12	210	10	220	14.68%	4.55%	5.24%	11
43	4	4	0	28	702	22	724	28.94%	3.04%	3.99%	18
44	4	5	1	33	556	13	569	20.25%	2.28%	5.76%	22
45	4	6	1	22	407	11	418	20.99%	2.63%	5.16%	19
46	4	7	0	27	604	13	617	35.02%	2.11%	4.47%	21
47	4	8	0	25	468	11	479	21.77%	2.30%	5.34%	31
48	**Ward 4 Subtotal**		**12**	**187**	**3,810**	**130**	**3,940**	**23.06%**	**3.30%**	**4.59%**	**179**

Are these data tidy?

- No!
- Most of the rows represent observations for single precincts, but some give ward and city-wide totals

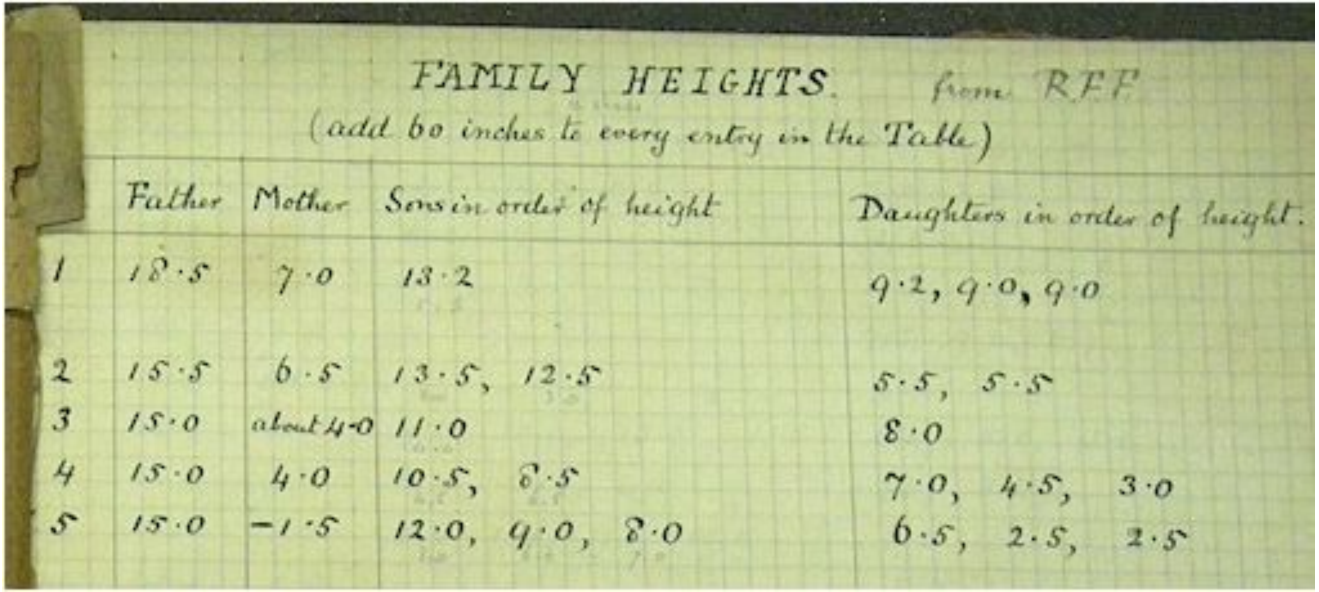
Tidy Data?



City of Minneapolis
General Election November 5, 2013
SUMMARY

	A	B	C	D	E	F	G	H	I	J	K	L
4			Registered Voters at 7am	Voters Registering at Polls	Voters Registering by Absentee	Total Registrations	Voters at Polls	Absente e Voters	Total Ballots Cast	Total Turnout	Percentage Absentee	Spoiled Ballots
5	WARD 1		18,836	364	9	373	5,725	217	5,942	30.93%	3.65%	189
6	WARD 2		18,196	493	16	509	4,950	206	5,156	27.56%	4.00%	196
7	WARD 3		19,364	623	40	663	5,778	428	6,206	30.99%	6.90%	260
8	WARD 4		16,899	175	12	187	3,810	130	3,940	23.06%	3.30%	179
9	WARD 5		15,013	335	40	375	3,419	202	3,621	23.53%	5.58%	320
10	WARD 6		14,026	623	374	997	3,388	1,663	5,051	33.62%	32.92%	287
11	WARD 7		19,178	456	17	473	6,212	382	6,594	33.56%	5.79%	274
12	WARD 8		16,859	437	26	463	5,852	210	6,062	35.00%	3.46%	266
13	WARD 9		12,138	496	24	520	4,140	170	4,310	34.05%	3.94%	228
14	WARD 10		18,616	786	54	840	5,636	297	5,933	30.49%	5.01%	209
15	WARD 11		19,720	330	50	380	7,494	306	7,800	38.81%	3.92%	256
16	WARD 12		21,660	413	35	448	8,428	314	8,742	39.54%	3.59%	285
17	WARD 13		22,846	395	11	406	10,313	429	10,742	46.20%	3.99%	409

Francis Galton



FAMILY HEIGHTS. from RFF

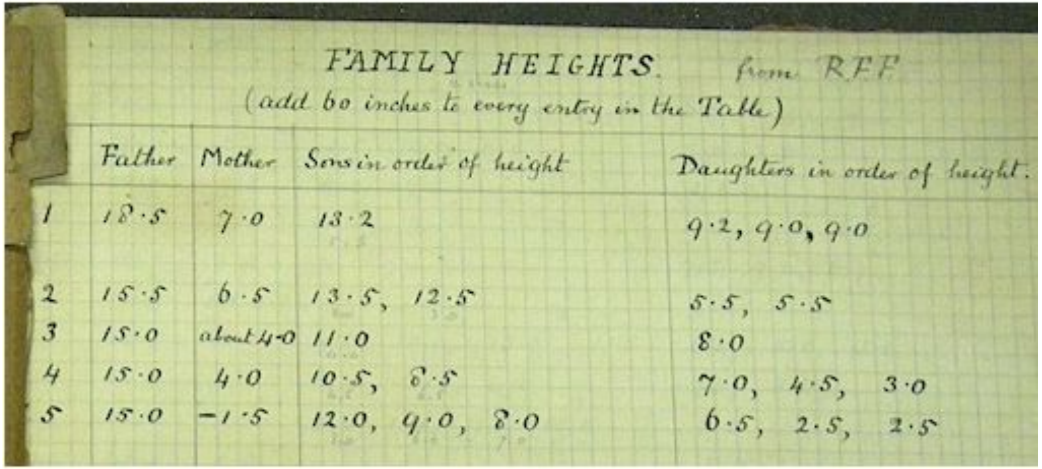
(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5

Image source

Tidy Data?

- What are the variables?
- What are the observations?
- Were Francis Galton’s data tidy?



FAMILY HEIGHTS from RFF

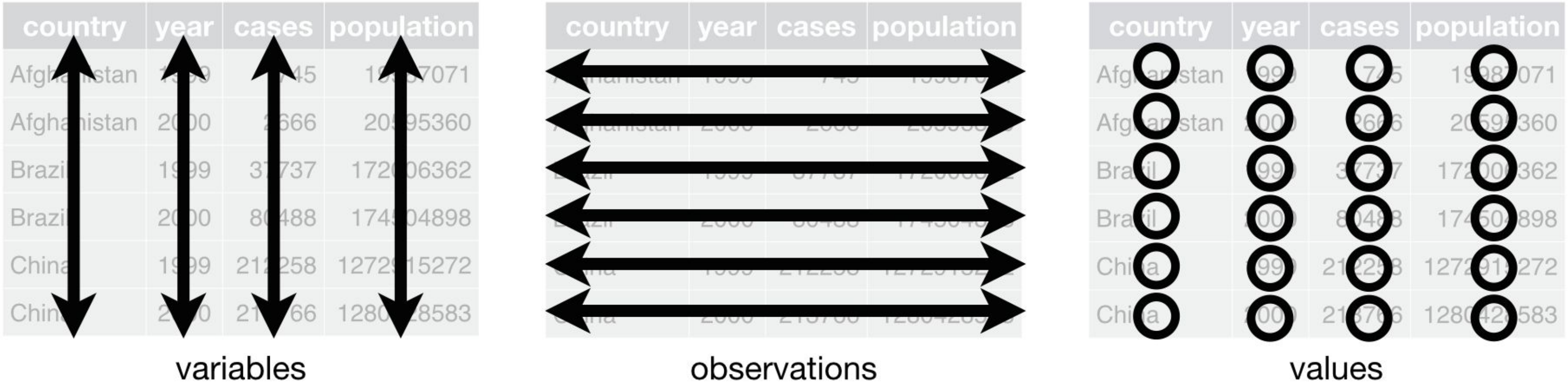
(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5

Image source

tidyr

Tidy Data



variables

observations

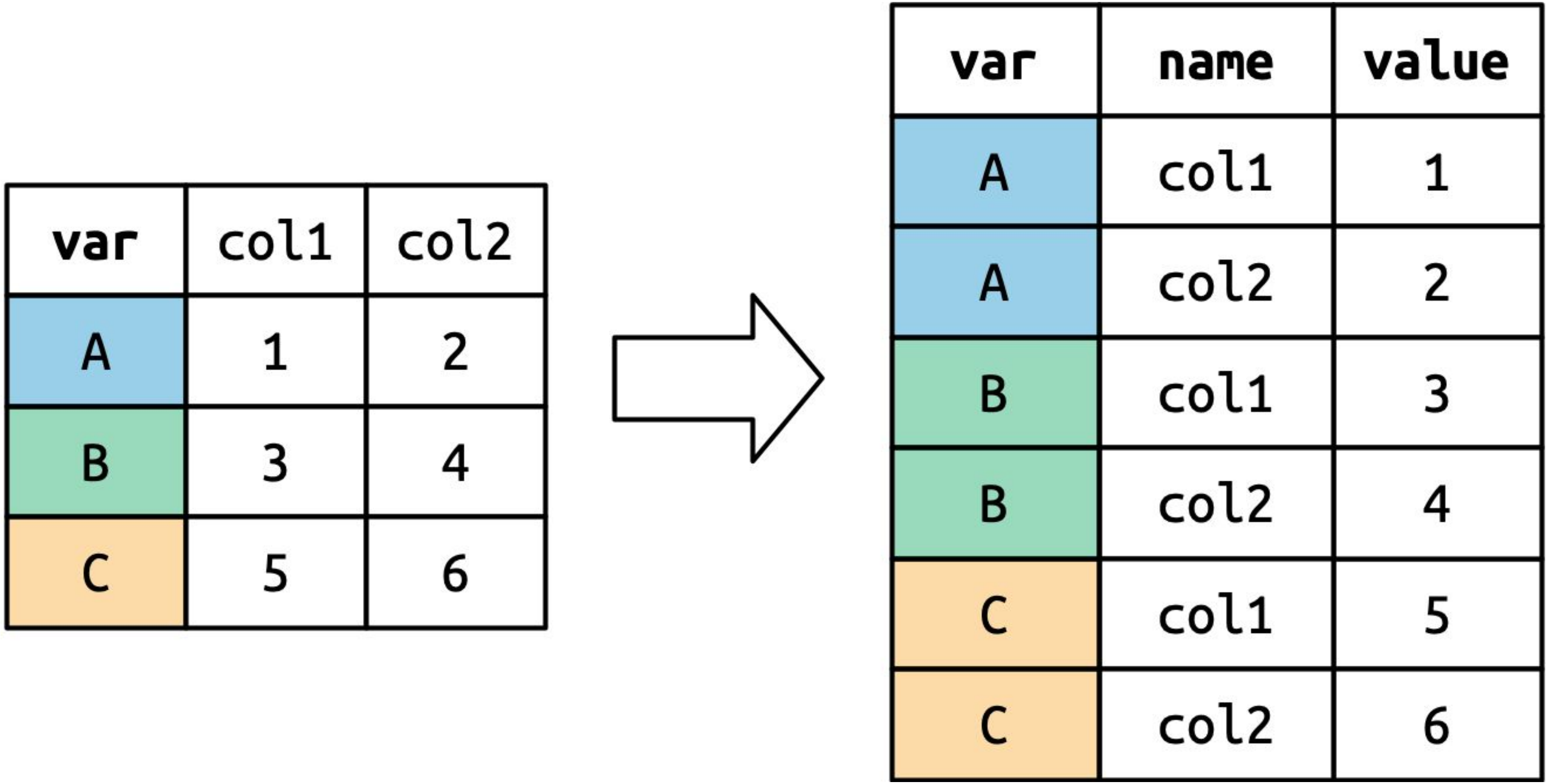
values

Image source

Some useful functions

- `gather`: Converts data from wide form to long form
- `spread`: Complement of gather (converts to wide form)
- `separate`: Splits a single variable into two
- `unite`: Complement of separate

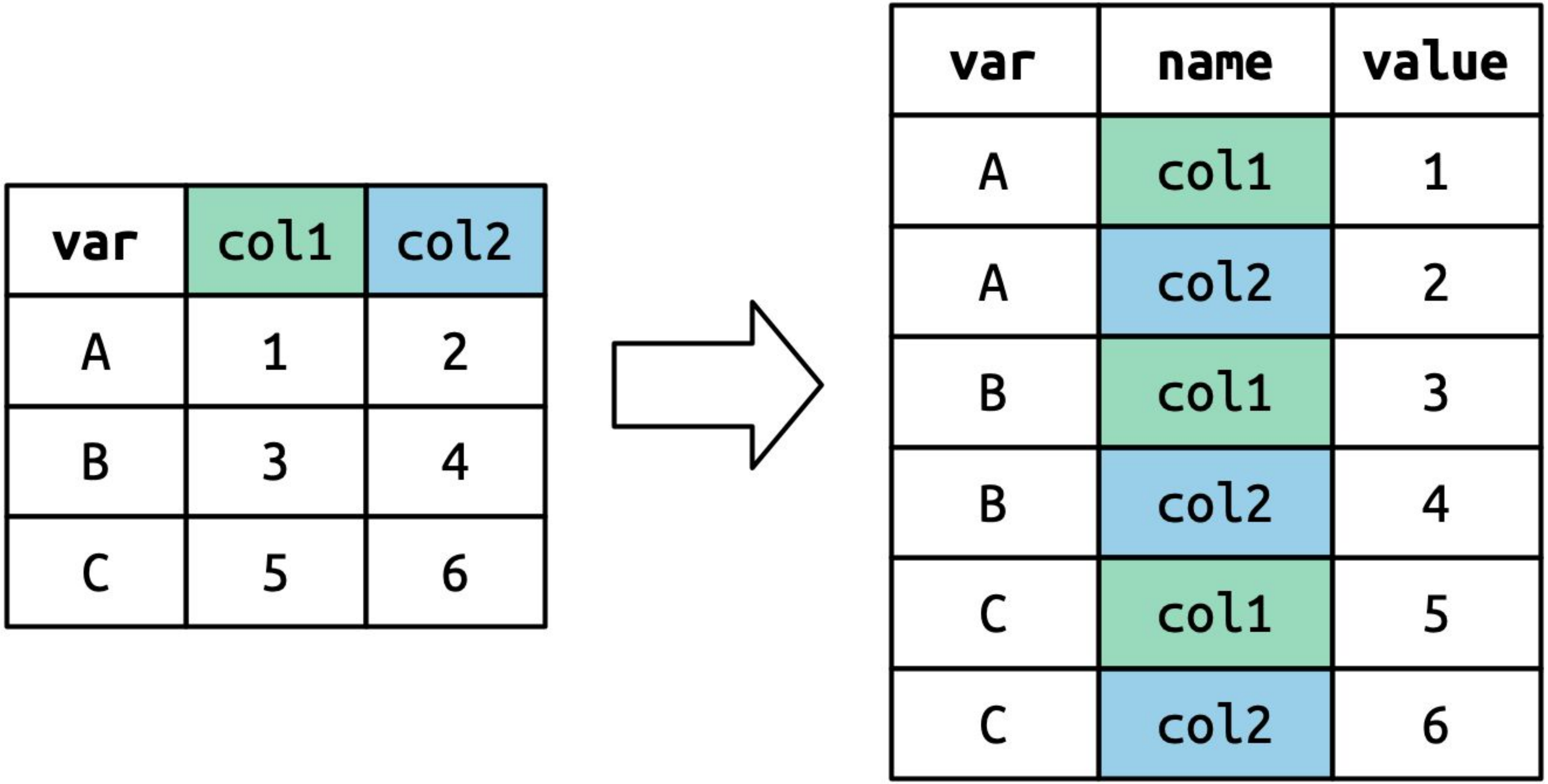
Gather: Observations



var	col1	col2
A	1	2
B	3	4
C	5	6

var	name	value
A	col1	1
A	col2	2
B	col1	3
B	col2	4
C	col1	5
C	col2	6

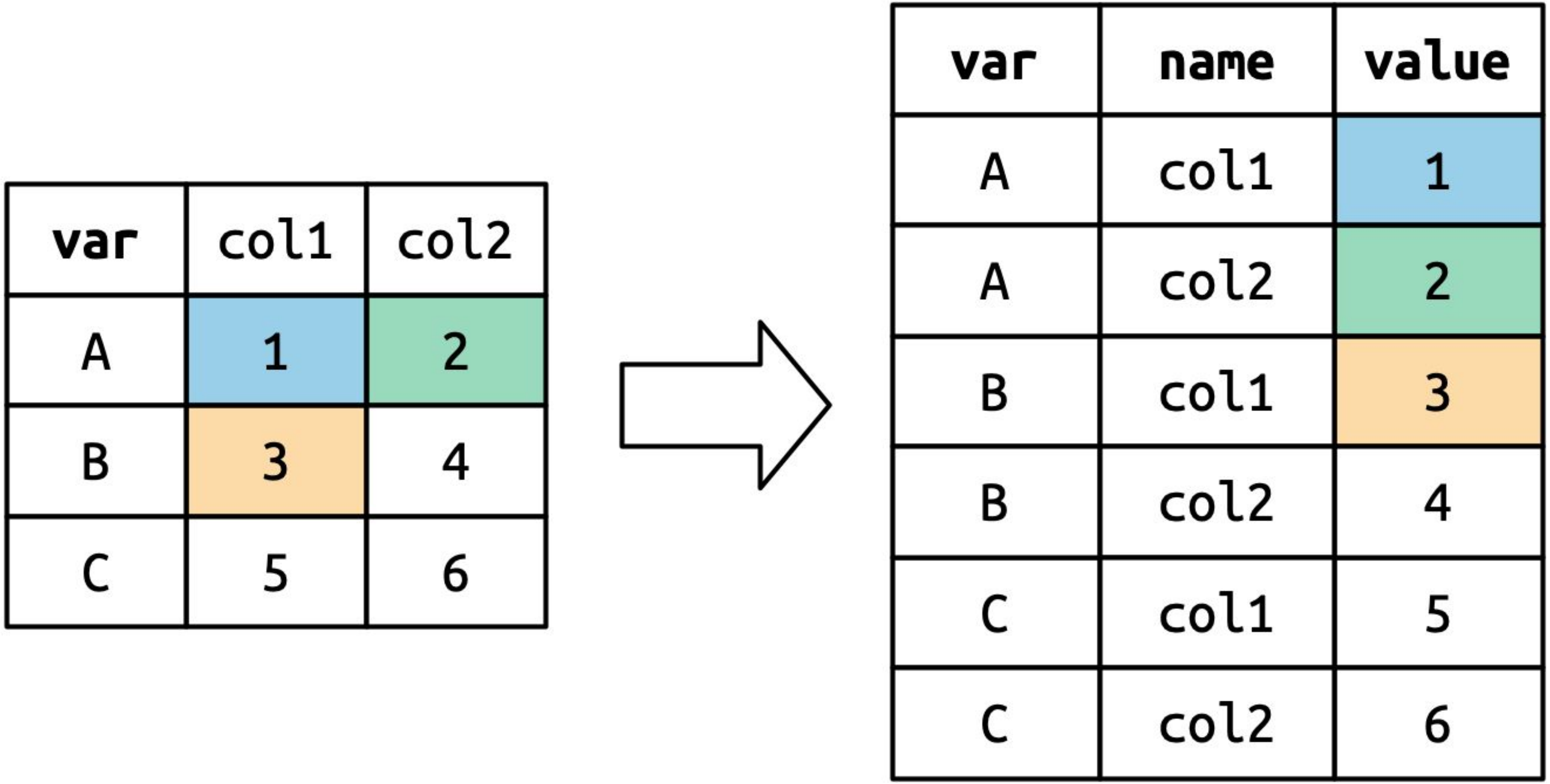
Gather: Variables



var	col1	col2
A	1	2
B	3	4
C	5	6

var	name	value
A	col1	1
A	col2	2
B	col1	3
B	col2	4
C	col1	5
C	col2	6

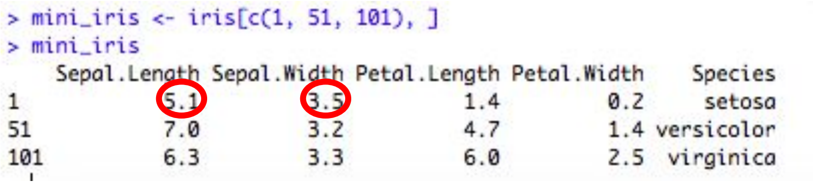
Gather: Values



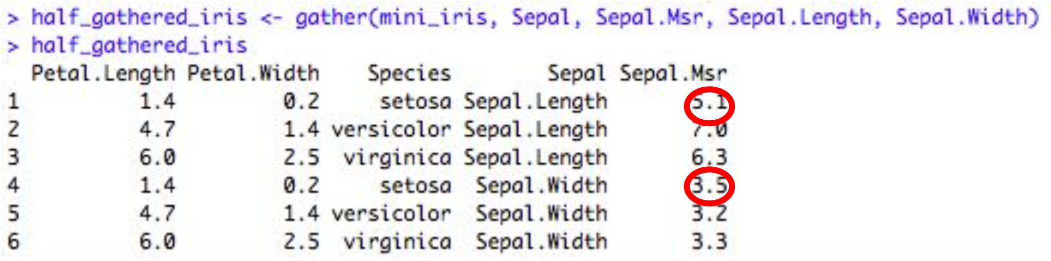
var	col1	col2
A	1	2
B	3	4
C	5	6

var	name	value
A	col1	1
A	col2	2
B	col1	3
B	col2	4
C	col1	5
C	col2	6

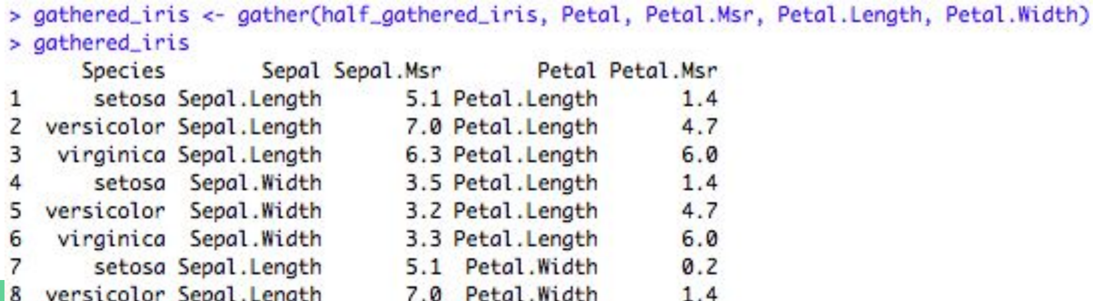
Gather



```
> mini_iris <- iris[c(1, 51, 101), ]
> mini_iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
51           7.0         3.2          4.7         1.4 versicolor
101          6.3         3.3          6.0         2.5  virginica
```



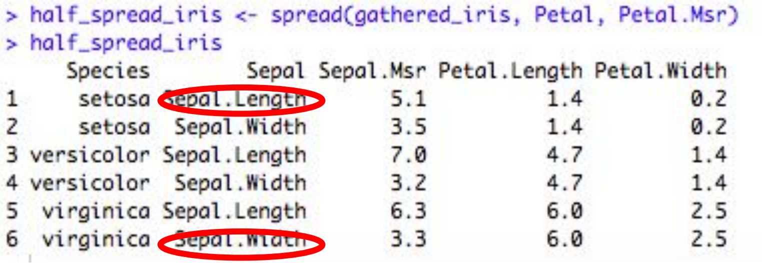
```
> half_gathered_iris <- gather(mini_iris, Sepal, Sepal.Msr, Sepal.Length, Sepal.Width)
> half_gathered_iris
  Petal.Length Petal.Width    Species        Sepal Sepal.Msr
1          1.4         0.2     setosa Sepal.Length       5.1
2          4.7         1.4 versicolor Sepal.Length       7.0
3          6.0         2.5  virginica Sepal.Length       6.3
4          1.4         0.2     setosa  Sepal.Width       3.5
5          4.7         1.4 versicolor  Sepal.Width       3.2
6          6.0         2.5  virginica  Sepal.Width       3.3
```



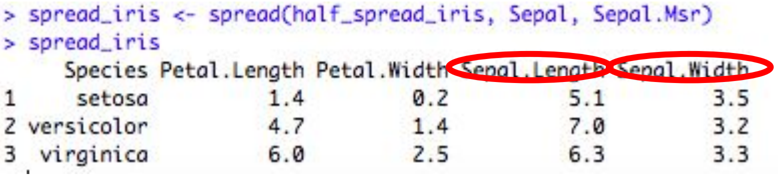
```
> gathered_iris <- gather(half_gathered_iris, Petal, Petal.Msr, Petal.Length, Petal.Width)
> gathered_iris
      Species        Sepal Sepal.Msr        Petal Petal.Msr
1      setosa Sepal.Length       5.1 Petal.Length       1.4
2  versicolor Sepal.Length       7.0 Petal.Length       4.7
3   virginica Sepal.Length       6.3 Petal.Length       6.0
4      setosa  Sepal.Width       3.5 Petal.Length       1.4
5  versicolor  Sepal.Width       3.2 Petal.Length       4.7
6   virginica  Sepal.Width       3.3 Petal.Length       6.0
7      setosa Sepal.Length       5.1  Petal.Width       0.2
8  versicolor Sepal.Length       7.0  Petal.Width       1.4
```

Spread

- Undoes the work of gather
- Converts from data from long form to wide form



```
> half_spread_iris <- spread(gathered_iris, Petal, Petal.Msr)
> half_spread_iris
     Species        Sepal Sepal.Msr Petal.Length Petal.Width
1     setosa Sepal.Length       5.1          1.4         0.2
2     setosa  Sepal.Width       3.5          1.4         0.2
3 versicolor Sepal.Length       7.0          4.7         1.4
4 versicolor  Sepal.Width       3.2          4.7         1.4
5  virginica Sepal.Length       6.3          6.0         2.5
6  virginica  Sepal.Width       3.3          6.0         2.5
```



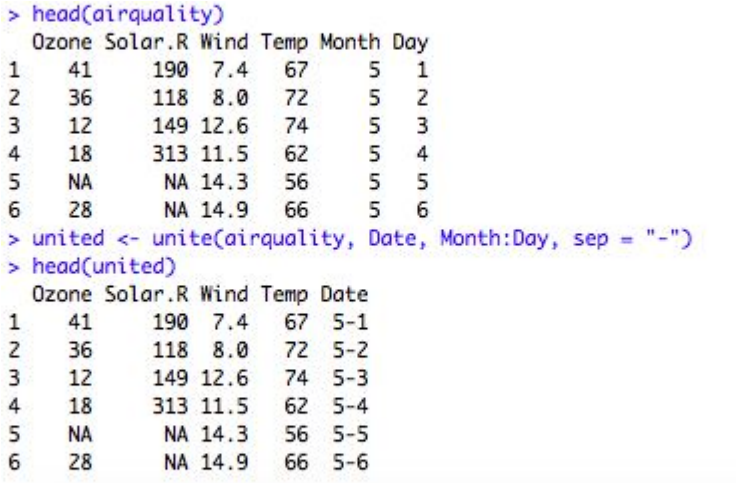
```
> spread_iris <- spread(half_spread_iris, Sepal, Sepal.Msr)
> spread_iris
     Species Petal.Length Petal.Width Sepal.Length Sepal.Width
1     setosa          1.4         0.2          5.1         3.5
2 versicolor          4.7         1.4          7.0         3.2
3  virginica          6.0         2.5          6.3         3.3
```

Separate

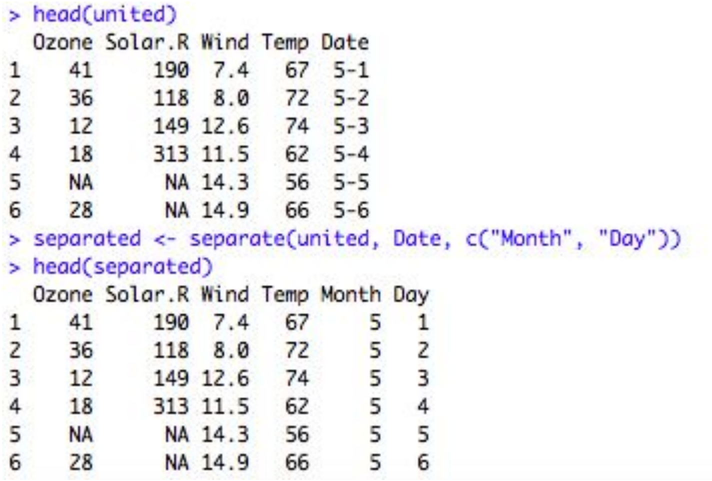
- Breaks up compound values into pieces
- Especially useful for breaking down date data

Unite

- Undoes the work of separate
- Concatenates multiple variables into one



```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
> united <- unite(airquality, Date, Month:Day, sep = "-")
> head(united)
  Ozone Solar.R Wind Temp Date
1    41     190  7.4   67  5-1
2    36     118  8.0   72  5-2
3    12     149 12.6   74  5-3
4    18     313 11.5   62  5-4
5    NA      NA 14.3   56  5-5
6    28      NA 14.9   66  5-6
```



```
> head(united)
  Ozone Solar.R Wind Temp Date
1    41     190  7.4   67  5-1
2    36     118  8.0   72  5-2
3    12     149 12.6   74  5-3
4    18     313 11.5   62  5-4
5    NA      NA 14.3   56  5-5
6    28      NA 14.9   66  5-6
> separated <- separate(united, Date, c("Month", "Day"))
> head(separated)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```