Stat645

Data structure & cleaning

Hadley Wickham

“Happy families are all alike; every unhappy family is unhappy in its own way.”
—Leo Tolstoy

"Clean datasets are all alike; every messy dataset is messy in its own way."

—Hadley Wickham

Rectangular data

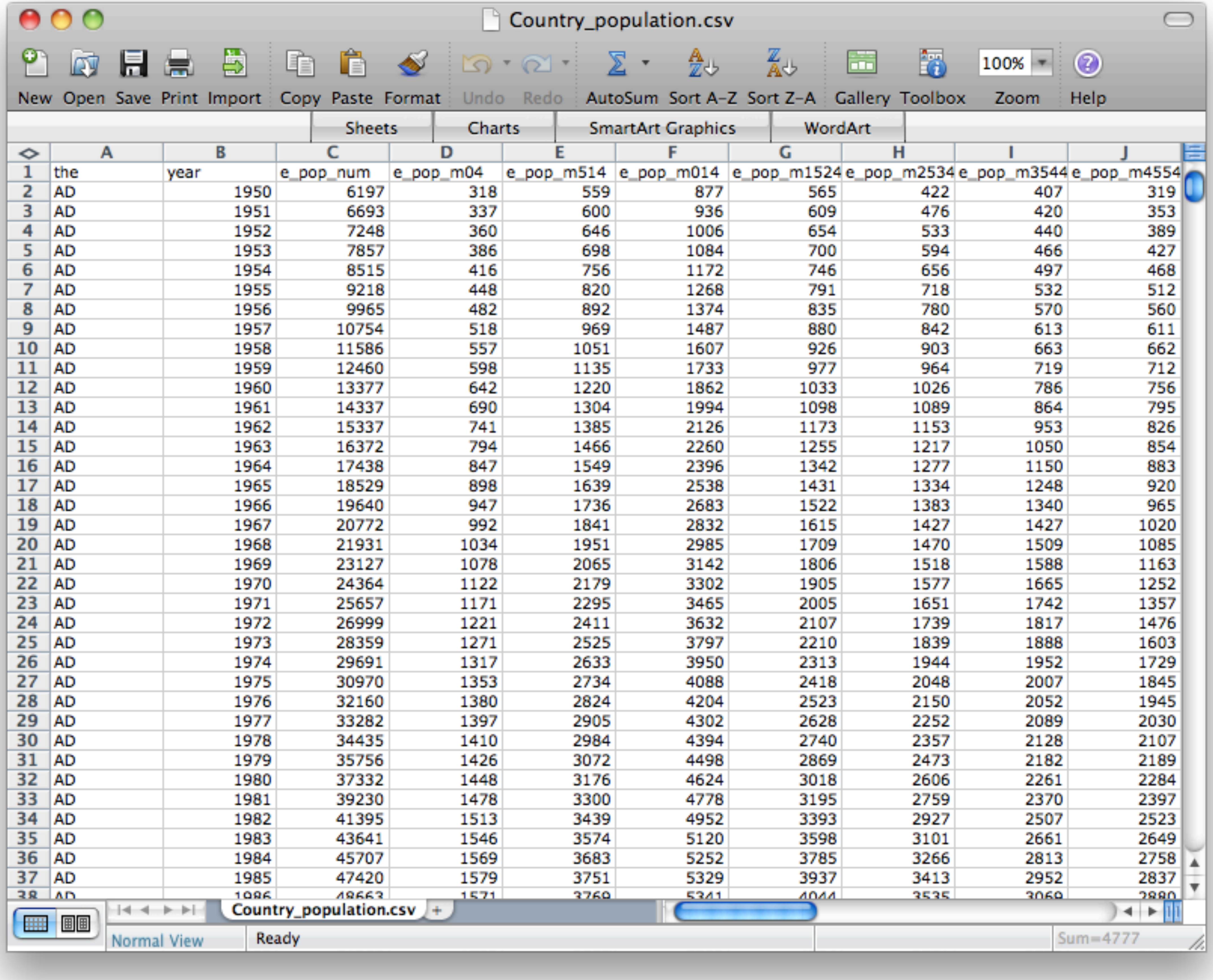
Pretty much all of the data we deal with is rectangular: has columns and rows.

Statistically, we want variables and observations.

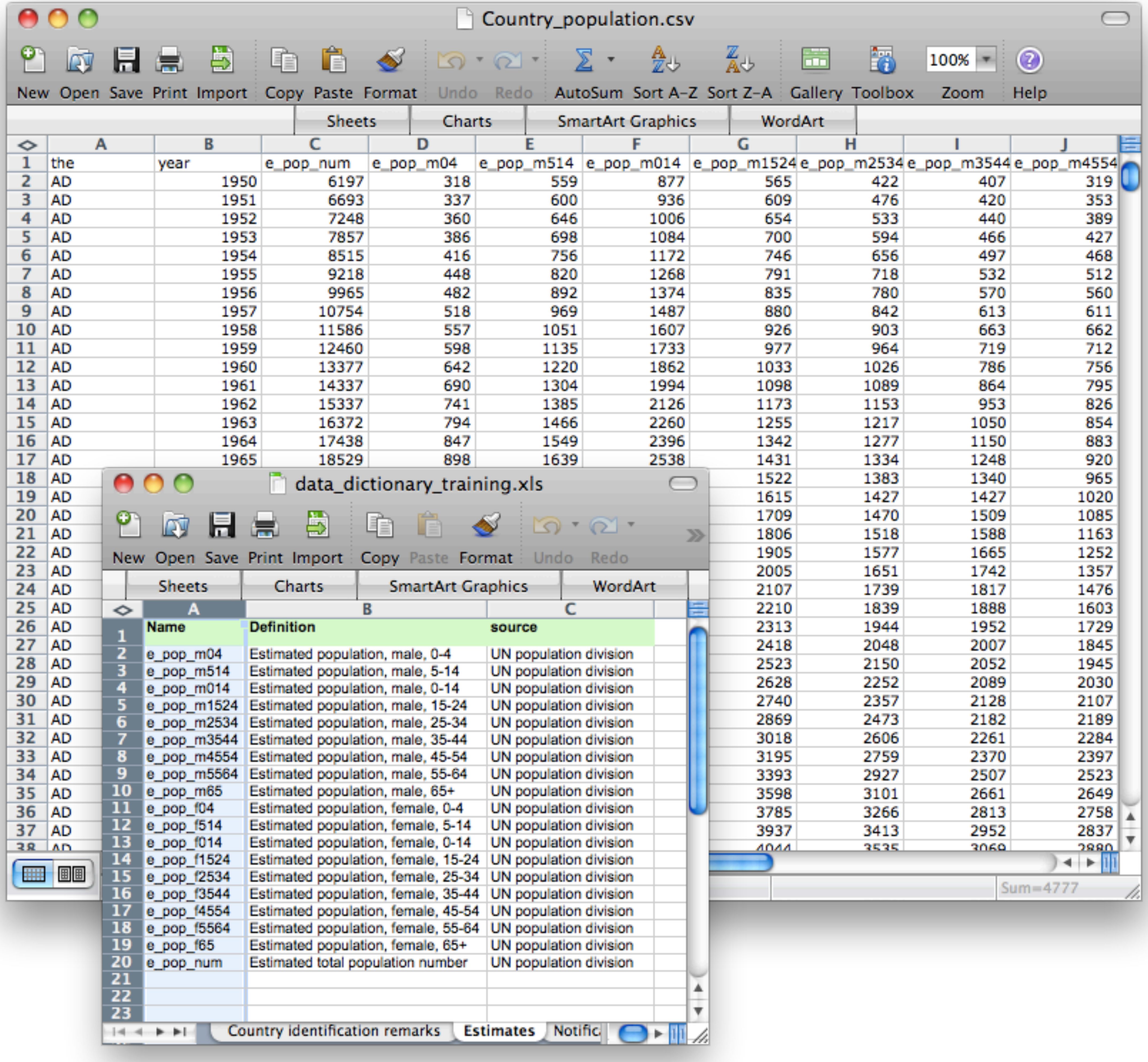
Variables go in columns, observations go in rows (with occasional exceptions)

Income Level by Religious Tradition

	Less than $30,000	$30,000-$49,999	$50,000-$74,999	$75,000-$99,999	$100,000+
	%	%	%	%	%
Total Population	31	22	17	13	18
Total Protestants	32	23	17	12	15
Members of Evangelical Prot. Churches	34	24	18	11	13
Members of Mainline Protestant Churches	25	21	18	15	21
Members of Hist. Black Prot. Churches	47	26	12	7	8
Catholic	31	20	16	14	19
Mormon	26	21	22	16	16
Church of Jesus Christ of Latter-day Saints	26	21	22	16	15
Jehovah's Witness	42	23	17	9	9
Orthodox	20	24	16	13	28
Greek Orthodox	17	22	18	13	30
Other Christian	29	21	13	13	23
Jewish	14	11	17	12	46
Reform	11	8	14	12	55
Conservative	12	14	17	14	43
Muslim*	35	24	15	10	16
Buddhist	25	19	17	17	22
Hindu	9	10	15	22	43
Other Faiths	28	25	16	13	18
Unitarian and Other Liberal Faiths	19	25	16	13	26
New Age	39	23	17	12	9
Unaffiliated	29	23	16	13	19
Atheist	21	20	16	15	28
Agnostic	18	22	19	16	25
Secular Unaffiliated	25	24	17	13	21
Religious Unaffiliated	40	24	15	10	12

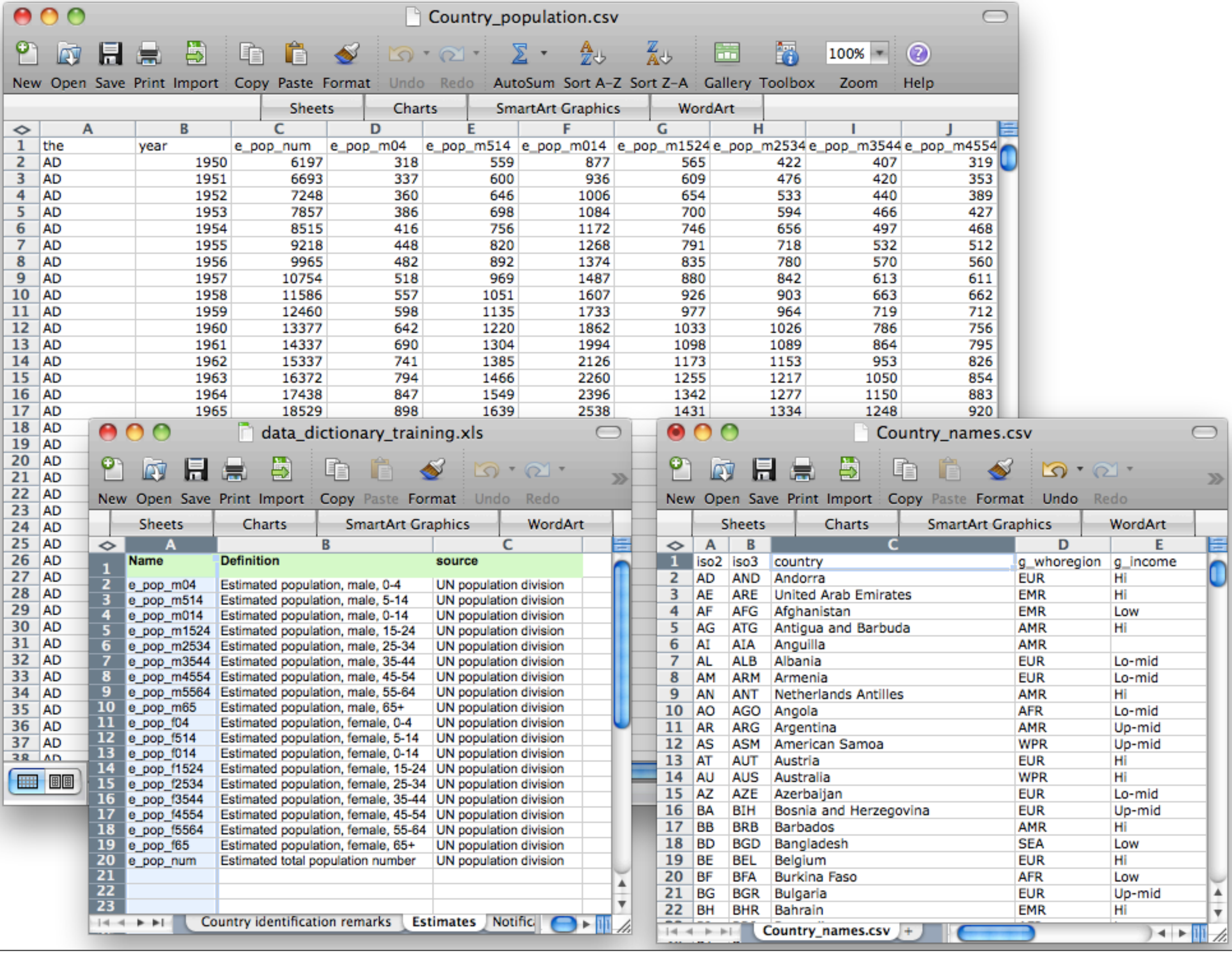


	A	B	C	D	E	F	G	H	I	J
1	the	year	e_pop_num	e_pop_m04	e_pop_m514	e_pop_m014	e_pop_m1524	e_pop_m2534	e_pop_m3544	e_pop_m4554
2	AD	1950	6197	318	559	877	565	422	407	319
3	AD	1951	6693	337	600	936	609	476	420	353
4	AD	1952	7248	360	646	1006	654	533	440	389
5	AD	1953	7857	386	698	1084	700	594	466	427
6	AD	1954	8515	416	756	1172	746	656	497	468
7	AD	1955	9218	448	820	1268	791	718	532	512
8	AD	1956	9965	482	892	1374	835	780	570	560
9	AD	1957	10754	518	969	1487	880	842	613	611
10	AD	1958	11586	557	1051	1607	926	903	663	662
11	AD	1959	12460	598	1135	1733	977	964	719	712
12	AD	1960	13377	642	1220	1862	1033	1026	786	756
13	AD	1961	14337	690	1304	1994	1098	1089	864	795
14	AD	1962	15337	741	1385	2126	1173	1153	953	826
15	AD	1963	16372	794	1466	2260	1255	1217	1050	854
16	AD	1964	17438	847	1549	2396	1342	1277	1150	883
17	AD	1965	18529	898	1639	2538	1431	1334	1248	920
18	AD	1966	19640	947	1736	2683	1522	1383	1340	965
19	AD	1967	20772	992	1841	2832	1615	1427	1427	1020
20	AD	1968	21931	1034	1951	2985	1709	1470	1509	1085
21	AD	1969	23127	1078	2065	3142	1806	1518	1588	1163
22	AD	1970	24364	1122	2179	3302	1905	1577	1665	1252
23	AD	1971	25657	1171	2295	3465	2005	1651	1742	1357
24	AD	1972	26999	1221	2411	3632	2107	1739	1817	1476
25	AD	1973	28359	1271	2525	3797	2210	1839	1888	1603
26	AD	1974	29691	1317	2633	3950	2313	1944	1952	1729
27	AD	1975	30970	1353	2734	4088	2418	2048	2007	1845
28	AD	1976	32160	1380	2824	4204	2523	2150	2052	1945
29	AD	1977	33282	1397	2905	4302	2628	2252	2089	2030
30	AD	1978	34435	1410	2984	4394	2740	2357	2128	2107
31	AD	1979	35756	1426	3072	4498	2869	2473	2182	2189
32	AD	1980	37332	1448	3176	4624	3018	2606	2261	2284
33	AD	1981	39230	1478	3300	4778	3195	2759	2370	2397
34	AD	1982	41395	1513	3439	4952	3393	2927	2507	2523
35	AD	1983	43641	1546	3574	5120	3598	3101	2661	2649
36	AD	1984	45707	1569	3683	5252	3785	3266	2813	2758
37	AD	1985	47420	1579	3751	5329	3937	3413	2952	2837
38	AD	1986	48663	1571	3769	5341	4044	3535	3069	2880



	A	B	C	D	E	F	G	H	I	J
1	the	year	e_pop_num	e_pop_m04	e_pop_m514	e_pop_m014	e_pop_m1524	e_pop_m2534	e_pop_m3544	e_pop_m4554
2	AD	1950	6197	318	559	877	565	422	407	319
3	AD	1951	6693	337	600	936	609	476	420	353
4	AD	1952	7248	360	646	1006	654	533	440	389
5	AD	1953	7857	386	698	1084	700	594	466	427
6	AD	1954	8515	416	756	1172	746	656	497	468
7	AD	1955	9218	448	820	1268	791	718	532	512
8	AD	1956	9965	482	892	1374	835	780	570	560
9	AD	1957	10754	518	969	1487	880	842	613	611
10	AD	1958	11586	557	1051	1607	926	903	663	662
11	AD	1959	12460	598	1135	1733	977	964	719	712
12	AD	1960	13377	642	1220	1862	1033	1026	786	756
13	AD	1961	14337	690	1304	1994	1098	1089	864	795
14	AD	1962	15337	741	1385	2126	1173	1153	953	826
15	AD	1963	16372	794	1466	2260	1255	1217	1050	854
16	AD	1964	17438	847	1549	2396	1342	1277	1150	883
17	AD	1965	18529	898	1639	2538	1431	1334	1248	920
18	AD						1522	1383	1340	965
19	AD						1615	1427	1427	1020
20	AD						1709	1470	1509	1085
21	AD						1806	1518	1588	1163
22	AD						1905	1577	1665	1252
23	AD						2005	1651	1742	1357
24	AD						2107	1739	1817	1476
25	AD						2210	1839	1888	1603
26	AD						2313	1944	1952	1729
27	AD						2418	2048	2007	1845
28	AD						2523	2150	2052	1945
29	AD						2628	2252	2089	2030
30	AD						2740	2357	2128	2107
31	AD						2869	2473	2182	2189
32	AD						3018	2606	2261	2284
33	AD						3195	2759	2370	2397
34	AD						3393	2927	2507	2523
35	AD						3598	3101	2661	2649
36	AD						3785	3266	2813	2758
37	AD						3937	3413	2952	2837
38	AD						4044	3535	3069	2880

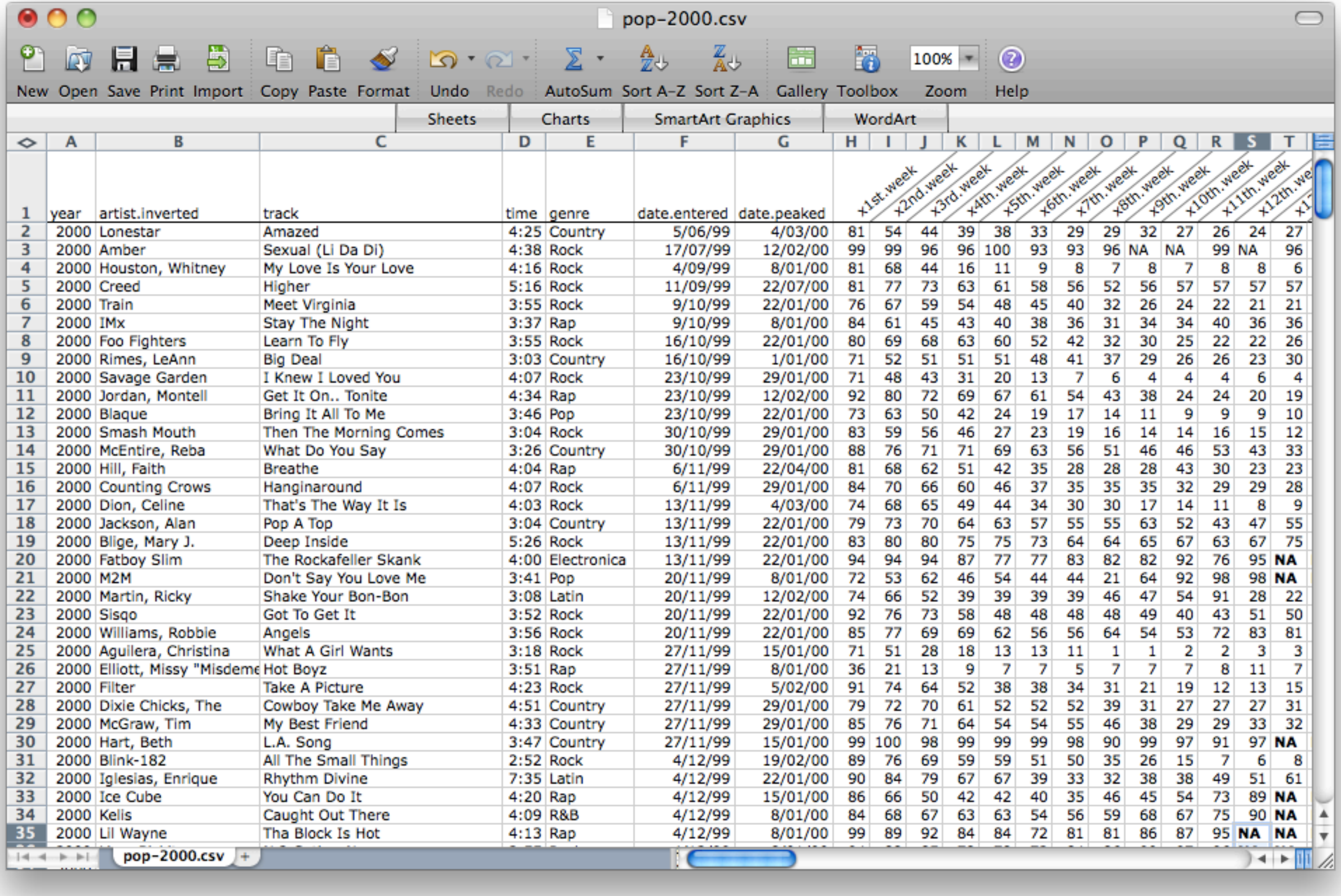
	A	B	C
1	**Name**	**Definition**	**source**
2	e_pop_m04	Estimated population, male, 0-4	UN population division
3	e_pop_m514	Estimated population, male, 5-14	UN population division
4	e_pop_m014	Estimated population, male, 0-14	UN population division
5	e_pop_m1524	Estimated population, male, 15-24	UN population division
6	e_pop_m2534	Estimated population, male, 25-34	UN population division
7	e_pop_m3544	Estimated population, male, 35-44	UN population division
8	e_pop_m4554	Estimated population, male, 45-54	UN population division
9	e_pop_m5564	Estimated population, male, 55-64	UN population division
10	e_pop_m65	Estimated population, male, 65+	UN population division
11	e_pop_f04	Estimated population, female, 0-4	UN population division
12	e_pop_f514	Estimated population, female, 5-14	UN population division
13	e_pop_f014	Estimated population, female, 0-14	UN population division
14	e_pop_f1524	Estimated population, female, 15-24	UN population division
15	e_pop_f2534	Estimated population, female, 25-34	UN population division
16	e_pop_f3544	Estimated population, female, 35-44	UN population division
17	e_pop_f4554	Estimated population, female, 45-54	UN population division
18	e_pop_f5564	Estimated population, female, 55-64	UN population division
19	e_pop_f65	Estimated population, female, 65+	UN population division
20	e_pop_num	Estimated total population number	UN population division
21			
22			
23			



	A	B	C	D	E	F	G	H	I	J
1	the	year	e_pop_num	e_pop_m04	e_pop_m514	e_pop_m014	e_pop_m1524	e_pop_m2534	e_pop_m3544	e_pop_m4554
2	AD	1950	6197	318	559	877	565	422	407	319
3	AD	1951	6693	337	600	936	609	476	420	353
4	AD	1952	7248	360	646	1006	654	533	440	389
5	AD	1953	7857	386	698	1084	700	594	466	427
6	AD	1954	8515	416	756	1172	746	656	497	468
7	AD	1955	9218	448	820	1268	791	718	532	512
8	AD	1956	9965	482	892	1374	835	780	570	560
9	AD	1957	10754	518	969	1487	880	842	613	611
10	AD	1958	11586	557	1051	1607	926	903	663	662
11	AD	1959	12460	598	1135	1733	977	964	719	712
12	AD	1960	13377	642	1220	1862	1033	1026	786	756
13	AD	1961	14337	690	1304	1994	1098	1089	864	795
14	AD	1962	15337	741	1385	2126	1173	1153	953	826
15	AD	1963	16372	794	1466	2260	1255	1217	1050	854
16	AD	1964	17438	847	1549	2396	1342	1277	1150	883
17	AD	1965	18529	898	1639	2538	1431	1334	1248	920
18	AD									
19	AD									
20	AD									
21	AD									
22	AD									
23	AD									
24	AD									
25	AD									
26	AD									
27	AD									
28	AD									
29	AD									
30	AD									
31	AD									
32	AD									
33	AD									
34	AD									
35	AD									
36	AD									
37	AD									

	A	B	C
1	**Name**	**Definition**	**source**
2	e_pop_m04	Estimated population, male, 0-4	UN population division
3	e_pop_m514	Estimated population, male, 5-14	UN population division
4	e_pop_m014	Estimated population, male, 0-14	UN population division
5	e_pop_m1524	Estimated population, male, 15-24	UN population division
6	e_pop_m2534	Estimated population, male, 25-34	UN population division
7	e_pop_m3544	Estimated population, male, 35-44	UN population division
8	e_pop_m4554	Estimated population, male, 45-54	UN population division
9	e_pop_m5564	Estimated population, male, 55-64	UN population division
10	e_pop_m65	Estimated population, male, 65+	UN population division
11	e_pop_f04	Estimated population, female, 0-4	UN population division
12	e_pop_f514	Estimated population, female, 5-14	UN population division
13	e_pop_f014	Estimated population, female, 0-14	UN population division
14	e_pop_f1524	Estimated population, female, 15-24	UN population division
15	e_pop_f2534	Estimated population, female, 25-34	UN population division
16	e_pop_f3544	Estimated population, female, 35-44	UN population division
17	e_pop_f4554	Estimated population, female, 45-54	UN population division
18	e_pop_f5564	Estimated population, female, 55-64	UN population division
19	e_pop_f65	Estimated population, female, 65+	UN population division
20	e_pop_num	Estimated total population number	UN population division
21			
22			
23			

	A	B	C	D	E
1	iso2	iso3	country	g_whoregion	g_income
2	AD	AND	Andorra	EUR	HI
3	AE	ARE	United Arab Emirates	EMR	HI
4	AF	AFG	Afghanistan	EMR	Low
5	AG	ATG	Antigua and Barbuda	AMR	HI
6	AI	AIA	Anguilla	AMR	
7	AL	ALB	Albania	EUR	Lo-mid
8	AM	ARM	Armenia	EUR	Lo-mid
9	AN	ANT	Netherlands Antilles	AMR	HI
10	AO	AGO	Angola	AFR	Lo-mid
11	AR	ARG	Argentina	AMR	Up-mid
12	AS	ASM	American Samoa	WPR	Up-mid
13	AT	AUT	Austria	EUR	HI
14	AU	AUS	Australia	WPR	HI
15	AZ	AZE	Azerbaijan	EUR	Lo-mid
16	BA	BIH	Bosnia and Herzegovina	EUR	Up-mid
17	BB	BRB	Barbados	AMR	HI
18	BD	BGD	Bangladesh	SEA	Low
19	BE	BEL	Belgium	EUR	HI
20	BF	BFA	Burkina Faso	AFR	Low
21	BG	BGR	Bulgaria	EUR	Up-mid
22	BH	BHR	Bahrain	EMR	HI



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	year	artist.inverted	track	time	genre	date.entered	date.peaked	x1st.week	x2nd.week	x3rd.week	x4th.week	x5th.week	x6th.week	x7th.week	x8th.week	x9th.week	x10th.week	x11th.week	x12th.we	x1
2	2000	Lonestar	Amazed	4:25	Country	5/06/99	4/03/00	81	54	44	39	38	33	29	29	32	27	26	24	27
3	2000	Amber	Sexual (Li Da Di)	4:38	Rock	17/07/99	12/02/00	99	99	96	96	100	93	93	96	NA	NA	99	NA	96
4	2000	Houston, Whitney	My Love Is Your Love	4:16	Rock	4/09/99	8/01/00	81	68	44	16	11	9	8	7	8	7	8	8	6
5	2000	Creed	Higher	5:16	Rock	11/09/99	22/07/00	81	77	73	63	61	58	56	52	56	57	57	57	57
6	2000	Train	Meet Virginia	3:55	Rock	9/10/99	22/01/00	76	67	59	54	48	45	40	32	26	24	22	21	21
7	2000	IMx	Stay The Night	3:37	Rap	9/10/99	8/01/00	84	61	45	43	40	38	36	31	34	34	40	36	36
8	2000	Foo Fighters	Learn To Fly	3:55	Rock	16/10/99	22/01/00	80	69	68	63	60	52	42	32	30	25	22	22	26
9	2000	Rimes, LeAnn	Big Deal	3:03	Country	16/10/99	1/01/00	71	52	51	51	51	48	41	37	29	26	26	23	30
10	2000	Savage Garden	I Knew I Loved You	4:07	Rock	23/10/99	29/01/00	71	48	43	31	20	13	7	6	4	4	4	6	4
11	2000	Jordan, Montell	Get It On.. Tonite	4:34	Rap	23/10/99	12/02/00	92	80	72	69	67	61	54	43	38	24	24	20	19
12	2000	Blaque	Bring It All To Me	3:46	Pop	23/10/99	22/01/00	73	63	50	42	24	19	17	14	11	9	9	9	10
13	2000	Smash Mouth	Then The Morning Comes	3:04	Rock	30/10/99	29/01/00	83	59	56	46	27	23	19	16	14	14	16	15	12
14	2000	McEntire, Reba	What Do You Say	3:26	Country	30/10/99	29/01/00	88	76	71	71	69	63	56	51	46	46	53	43	33
15	2000	Hill, Faith	Breathe	4:04	Rap	6/11/99	22/04/00	81	68	62	51	42	35	28	28	28	43	30	23	23
16	2000	Counting Crows	Hanginaround	4:07	Rock	6/11/99	29/01/00	84	70	66	60	46	37	35	35	35	32	29	29	28
17	2000	Dion, Celine	That's The Way It Is	4:03	Rock	13/11/99	4/03/00	74	68	65	49	44	34	30	30	17	14	11	8	9
18	2000	Jackson, Alan	Pop A Top	3:04	Country	13/11/99	22/01/00	79	73	70	64	63	57	55	55	63	52	43	47	55
19	2000	Blige, Mary J.	Deep Inside	5:26	Rock	13/11/99	22/01/00	83	80	80	75	75	73	64	64	65	67	63	67	75
20	2000	Fatboy Slim	The Rockafeller Skank	4:00	Electronica	13/11/99	22/01/00	94	94	94	87	77	77	83	82	82	92	76	95	**NA**
21	2000	M2M	Don't Say You Love Me	3:41	Pop	20/11/99	8/01/00	72	53	62	46	54	44	44	21	64	92	98	98	**NA**
22	2000	Martin, Ricky	Shake Your Bon-Bon	3:08	Latin	20/11/99	12/02/00	74	66	52	39	39	39	39	46	47	54	91	28	22
23	2000	Sisqo	Got To Get It	3:52	Rock	20/11/99	22/01/00	92	76	73	58	48	48	48	48	49	40	43	51	50
24	2000	Williams, Robbie	Angels	3:56	Rock	20/11/99	22/01/00	85	77	69	69	62	56	56	64	54	53	72	83	81
25	2000	Aguilera, Christina	What A Girl Wants	3:18	Rock	27/11/99	15/01/00	71	51	28	18	13	13	11	1	1	2	2	3	3
26	2000	Elliott, Missy "Misdeme	Hot Boyz	3:51	Rap	27/11/99	8/01/00	36	21	13	9	7	7	5	7	7	7	8	11	7
27	2000	Filter	Take A Picture	4:23	Rock	27/11/99	5/02/00	91	74	64	52	38	38	34	31	21	19	12	13	15
28	2000	Dixie Chicks, The	Cowboy Take Me Away	4:51	Country	27/11/99	29/01/00	79	72	70	61	52	52	52	39	31	27	27	27	31
29	2000	McGraw, Tim	My Best Friend	4:33	Country	27/11/99	29/01/00	85	76	71	64	54	54	55	46	38	29	29	33	32
30	2000	Hart, Beth	L.A. Song	3:47	Country	27/11/99	15/01/00	99	100	98	99	99	99	98	90	99	97	91	97	**NA**
31	2000	Blink-182	All The Small Things	2:52	Rock	4/12/99	19/02/00	89	76	69	59	59	51	50	35	26	15	7	6	8
32	2000	Iglesias, Enrique	Rhythm Divine	7:35	Latin	4/12/99	22/01/00	90	84	79	67	67	39	33	32	38	38	49	51	61
33	2000	Ice Cube	You Can Do It	4:20	Rap	4/12/99	15/01/00	86	66	50	42	42	40	35	46	45	54	73	89	**NA**
34	2000	Kelis	Caught Out There	4:09	R&B	4/12/99	8/01/00	84	68	67	63	63	54	56	59	68	67	75	90	**NA**
35	2000	Lil Wayne	Tha Block Is Hot	4:13	Rap	4/12/99	8/01/00	99	89	92	84	84	72	81	81	86	87	95	**NA**	**NA**

Common problems

- One variable spread over multiple columns
- One column representing multiple variables
- Both together

Solution

- Identify variables
- Melt data to get fix variables spread over multiple columns
- Split apart columns that represent multiple variables (often with join or string operations)
- Convert back to long form, if necessary

Solution

- **Identify variables**
- Melt data to get fix variables spread over multiple columns
- Split apart columns that represent multiple variables (often with join or string operations)
- Convert back to long form, if necessary

What is a variable?

- "I know it when I see it"
- A variable is a class, not a value: sex is a variable, male and female are values
- Every value in a dataset is either a value or a variable name. Every value is associated with a variable.

```
MX000766800198903TMAX  270  G  270  G  279  G  275  G  284  G  260  G  225  G  215  G  220
MX000766800198904TMAX  260  G  256  G  251  G  265  G  262  G  230  G  230  G  255  G  269
MX000766800198904TMIN  118  G   90  G   96  G  104  G  102  G  110  G   80  G   85  G   88
MX000766800198905TMAX  294  G  250  G-9999     250  G  258  G  260  G  260  G  258  G  258
MX000766800198906TMAX  280  G  280  G  275  G  280  G  290  G-9999     300  G  335  G  315
MX000766800198907TMAX  245  G  212  G-9999     260  G  242  G  245  G  188  G  244  G  288
MX000766800198908TMAX  227  G-9999   -9999   -9999   -9999   -9999     239  G-9999   -9999
MX000766800198910TMAX-9999   -9999     256  G  257  G  246  G  260  G  255  G  236  G  218
MX000766800198911TMAX  225  G  235  G  232  G  228  G  247  G  263  G  260  G  265  G  260
MX000766800198912TMAX  222  G  218  G  211  G  168  G  175  G  195  G  220  G  200  G  117
MX000766800199001TMAX  237  G  210  G  224  G  222  G  232  G  245  G  240  G  240  G  225
MX000766800199002TMAX  245  G  222  G  262  G  250  G  232  G  236  G  230  G  250  G-9999
MX000766800199003TMAX  225  G  232  G-9999     248  G  250  G  250  G  262  G-9999   -9999
MX000766800199004TMAX  245  G  275  G  285  G-9999     249  G  275  G-9999     270  G  278
MX000766800199005TMAX  280  G  300  G  280  G  299  G  270  G  232  G  172  G  180  G  220
MX000766800199006TMAX  290  G-9999   -9999     282  G  262  G  275  G-9999     272  G  278
MX000766800199007TMAX  203  G  201  G-9999     235  G  249  G  250  G  217  G  230  G-9999
MX000766800199008TMAX  228  G-9999     245  G  250  G  260  G  250  G  200  G  160  G  240
MX000766800199009TMAX  248  G  255  G  220  G  235  G  240  G  223  G  245  G-9999     235
MX000766800199010TMAX  263  G  260  G  280  G  272  G-9999   -9999     227  G  275  G-9999
MX000766800199011TMAX  212  G  230  G  230  G  250  G  260  G  247  G  275  G  270  G  250
MX000766800199012TMAX  200  G  222  G  260  G  248  G  129  G  225  G  220  G  190  G  151
MX000766800199101TMAX  239  G  226  G  224  G  225  G  223  G  230  G  231  G  234  G  231
```

Day 1 2 3 ...

```
MX000766800198903TMAX  270  G  270  G  279  G  275  G  284  G  260  G  225  G  215  G  220
MX000766800198904TMAX  260  G  256  G  251  G  265  G  262  G  230  G  230  G  255  G  269
MX000766800198904TMIN  118  G   90  G   96  G  104  G  102  G  110  G   80  G   85  G   88
MX000766800198905TMAX  294  G  250  G-9999     250  G  258  G  260  G  260  G  258  G  258
MX000766800198906TMAX  280  G  280  G  275  G  280  G  290  G-9999     300  G  335  G  315
MX000766800198907TMAX  245  G  212  G-9999     260  G  242  G  245  G  188  G  244  G  288
MX000766800198908TMAX  227  G-9999   -9999   -9999   -9999   -9999     239  G-9999   -9999
MX000766800198910TMAX-9999   -9999     256  G  257  G  246  G  260  G  255  G  236  G  218
MX000766800198911TMAX  225  G  235  G  232  G  228  G  247  G  263  G  260  G  265  G  260
MX000766800198912TMAX  222  G  218  G  211  G  168  G  175  G  195  G  220  G  200  G  117
MX000766800199001TMAX  237  G  210  G  224  G  222  G  232  G  245  G  240  G  240  G  225
MX000766800199002TMAX  245  G  222  G  262  G  250  G  232  G  236  G  230  G  250  G-9999
MX000766800199003TMAX  225  G  232  G-9999     248  G  250  G  250  G  262  G-9999   -9999
MX000766800199004TMAX  245  G  275  G  285  G-9999     249  G  275  G-9999     270  G  278
MX000766800199005TMAX  280  G  300  G  280  G  299  G  270  G  232  G  172  G  180  G  220
MX000766800199006TMAX  290  G-9999   -9999     282  G  262  G  275  G-9999     272  G  278
MX000766800199007TMAX  203  G  201  G-9999     235  G  249  G  250  G  217  G  230  G-9999
MX000766800199008TMAX  228  G-9999     245  G  250  G  260  G  250  G  200  G  160  G  240
MX000766800199009TMAX  248  G  255  G  220  G  235  G  240  G  223  G  245  G-9999     235
MX000766800199010TMAX  263  G  260  G  280  G  272  G-9999   -9999     227  G  275  G-9999
MX000766800199011TMAX  212  G  230  G  230  G  250  G  260  G  247  G  275  G  270  G  250
MX000766800199012TMAX  200  G  222  G  260  G  248  G  129  G  225  G  220  G  190  G  151
MX000766800199101TMAX  239  G  226  G  224  G  225  G  223  G  230  G  231  G  234  G  231
```

Severe example

Your turn

Identify the variables in each of the first three examples.

Solution

- Identify variables
- **Melt data to get fix variables spread over multiple columns**
- Split apart columns that represent multiple variables (often with join or string operations)
- Convert back to long form, if necessary

```
# If you don't have reshape2 installed:
# install.packages("reshape2")

library(reshape2)
library(stringr)
options(stringsAsFactors = FALSE)

# Load -----------------------------------------------------------------------
note_raw <- read.csv("tb/TB_notification.csv")
note_raw$new_sp <- NULL

pop_raw <- read.csv("tb/Country_population.csv")
pop_raw$e_pop_num <- NULL
pop_raw <- subset(pop_raw, year < 2010)

# Melt -----------------------------------------------------------------------

note <- melt(note_raw, id = c("iso2", "year"), na.rm = TRUE)
names(note)[4] <- "cases"

pop <- melt(pop_raw, id = c("iso2", "year"), na.rm = TRUE)
names(pop)[4] <- "pop"
```

```
# If you don't have reshape2 installed:
# install.packages("reshape2")

library(reshape2)
library(stringr)
options(stringsAsFactors = FALSE)

# Load ------------------------------------------------------------------------
note_raw <- read.csv("tb/TB_notification.csv")
note_raw$new_sp <- NULL

pop_raw <- read.csv("tb/Country_population.csv")
pop_raw$e_pop_num <- NULL
pop_raw <- subset(

# Melt ----------                     ---------------------------------------------

note <- melt(note_raw, id = c("iso2", "year"), na.rm = TRUE)
names(note)[4] <- "cases"

pop <- melt(pop_raw, id = c("iso2", "year"), na.rm = TRUE)
names(pop)[4] <- "pop"
```

Columns that are already variables

Solution

- Identify variables
- Melt data to get fix variables spread over multiple columns
- **Split apart columns that represent multiple variables (often with join or string operations)**
- Convert back to long form, if necessary

```
# Break up variable into sex and age ----------------------------

note$variable <- str_replace(note$variable, "new_sp_", "")
pop$variable <- str_replace(pop$variable, "e_pop_", "")

note$sex <- str_sub(note$variable, 1, 1)
pop$sex <- str_sub(pop$variable, 1, 1)

ages <- c("04" = "0-4", "514" = "5-14", "014" = "0-14", "1524" =
"15-24", "2534" = "25-34", "3544" = "35-44", "4554" = "45-54",
"5564" = "55-64", "65"= "65+", "u" = NA)

pop$age <- factor(ages[str_sub(pop$variable, 2)], levels = ages)
note$age <- factor(ages[str_sub(note$variable, 2)], levels = ages)

pop$variable <- NULL
note$variable <- NULL
```

Your turn

Clean up the billboard data (pop-2000.csv) by following this same pattern.

```
# Remove values from songs that didn’t make it
# that long
dim(popm)
popm <- subset(popm, !is.na(value))
dim(popm)

# install.packages(”lubridate”)
library(lubridate)

# Calculate the actual date and focus on 2000
popm$date <- ymd(popm$date.entered) +
  weeks(popm$week - 1)
popm <- subset(popm, year(date) == 2000)
```

Your turn

Plot the data in an informative way!

For inspiration look at http://nyti.ms/mj-vis (partially constructed in R)

Hint: + scale_y_reverse()

Advanced reading

http://directlabels.r-forge.r-project.org

http://directlabels.r-forge.r-project.org/motivation.html

Solution

- Identify variables
- Melt data to get fix variables spread over multiple columns
- Split apart columns that represent multiple variables (often with join or string operations)
- **Convert back to long form, if necessary**

```
# None of these examples have needed it
# But if it did, you'd do something like

dcast(molten, ... + variable)
```

Next time

Read the “layered grammar of graphics” and write a one page response, following the guidelines on the website.

Focus on your reactions to the article, not the content.

Relational data

You may have noticed that some of the variables get repeated many many times. This is usually an indication that you have data on fundamentally different entities, that can’t be represented concisely in a single table.

This is known as relational data.