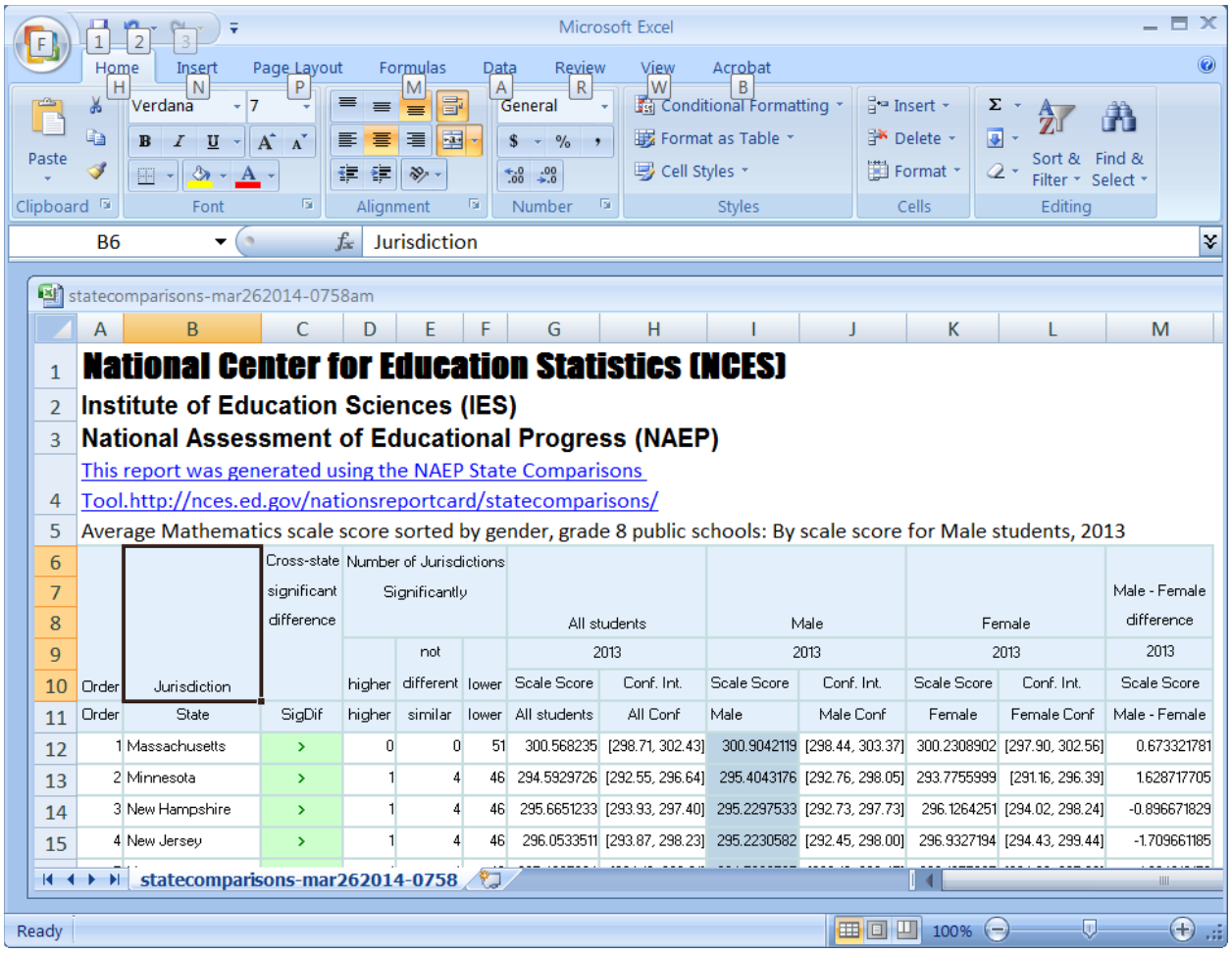
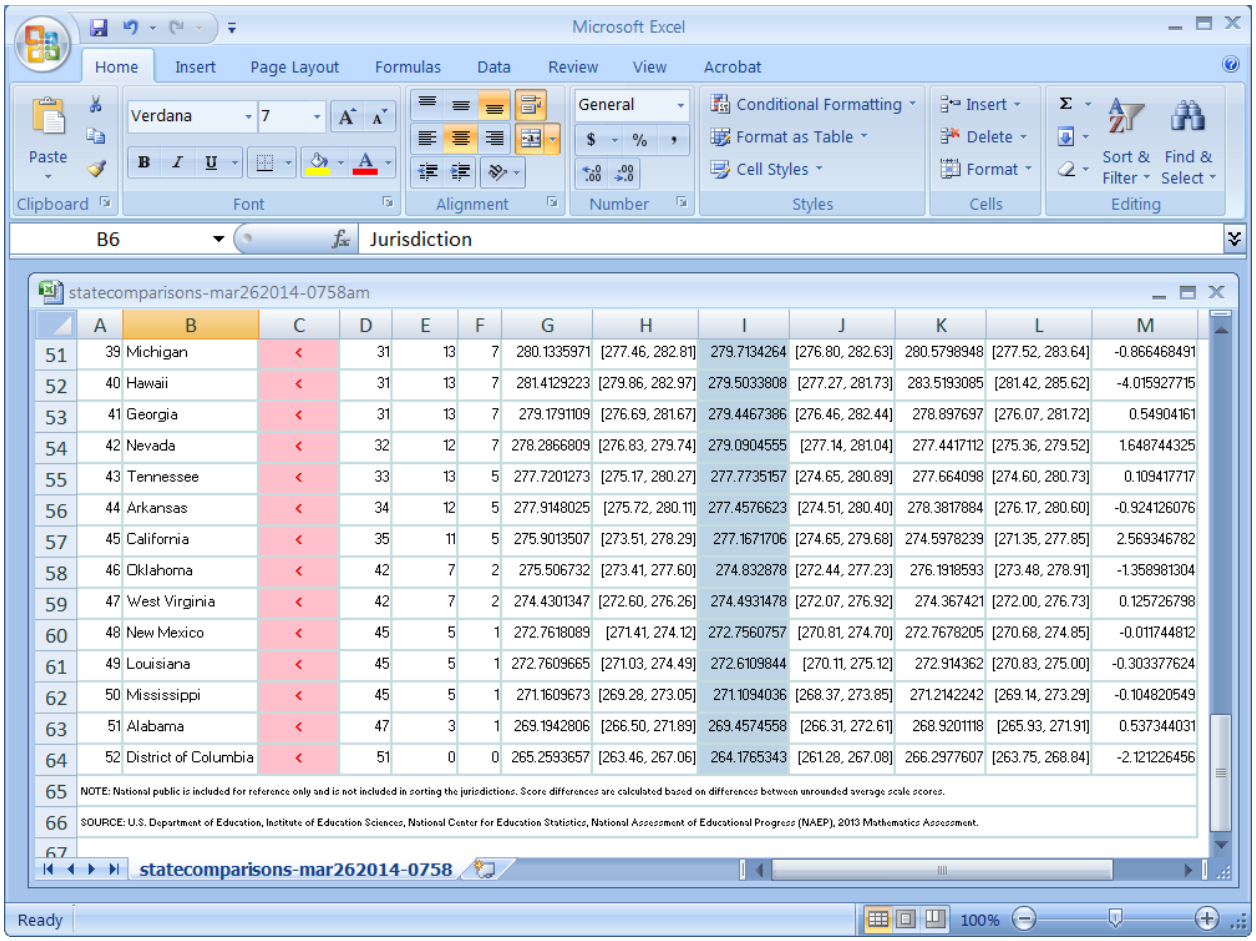
Preparing State Data For R and MicromapsST
A National Center for Education Statistics Example

The Federal agencies often provide data in Excel readable formats. The files often contain useful information and formatting that are not suitable for the R data.frames that we want to use in graphics production. For many US state micromaps we just need the state names (postal or fips codes will suffice), estimates and sometimes standard error and confidence intervals in separate columns. The example below is from National Center for Education Statistics (NCES) Excel file. This has

1) header and footer information that data.frames don't support,
2) Confidence intervals in single column, the values bound by brackets and separate by commas that date.frames don't support, and
3) Rows for non-state regions such as DoDEA and US National Public, the MicromapST doesn't support. Note: MicromapST expects the District of Columbia to be included as if it were a state. This is a federal standard.

If we were to process many such NCES files, it could be worth the effort to create R script that would do all the file preparation. Here we will use Excel to address 1) and 2) and prepare a comma delimited file for R to read. Then the R script just needs to remove the unwanted rows to produce the desired data.frame and can save the US National Public values for possible use as reference values. We also may want to further, easy revisions to the data.frame such modifying column labels





	A	B	C	D	E	F	G	H	I	J	K	L	M
51	39	Michigan	<	31	13	7	280.1335971	[277.46, 282.81]	279.7134264	[276.80, 282.63]	280.5798948	[277.52, 283.64]	-0.866468491
52	40	Hawaii	<	31	13	7	281.4129223	[279.86, 282.97]	279.5033808	[277.27, 281.73]	283.5193085	[281.42, 285.62]	-4.015927715
53	41	Georgia	<	31	13	7	279.1791109	[276.69, 281.67]	279.4467386	[276.46, 282.44]	278.897697	[276.07, 281.72]	0.54904161
54	42	Nevada	<	32	12	7	278.2866809	[276.83, 279.74]	279.0904555	[277.14, 281.04]	277.4417112	[275.36, 279.52]	1.648744325
55	43	Tennessee	<	33	13	5	277.7201273	[275.17, 280.27]	277.7735157	[274.65, 280.89]	277.664098	[274.60, 280.73]	0.109417717
56	44	Arkansas	<	34	12	5	277.9148025	[275.72, 280.11]	277.4576623	[274.51, 280.40]	278.3817884	[276.17, 280.60]	-0.924126076
57	45	California	<	35	11	5	275.9013507	[273.51, 278.29]	277.1671706	[274.65, 279.68]	274.5978239	[271.35, 277.85]	2.569346782
58	46	Oklahoma	<	42	7	2	275.506732	[273.41, 277.60]	274.832878	[272.44, 277.23]	276.1918593	[273.48, 278.91]	-1.358981304
59	47	West Virginia	<	42	7	2	274.4301347	[272.60, 276.26]	274.4931478	[272.07, 276.92]	274.367421	[272.00, 276.73]	0.125726798
60	48	New Mexico	<	45	5	1	272.7618089	[271.41, 274.12]	272.7560757	[270.81, 274.70]	272.7678205	[270.68, 274.85]	-0.011744812
61	49	Louisiana	<	45	5	1	272.7609665	[271.03, 274.49]	272.6109844	[270.11, 275.12]	272.914362	[270.83, 275.00]	-0.303377624
62	50	Mississippi	<	45	5	1	271.1609673	[269.28, 273.05]	271.1094036	[268.37, 273.85]	271.2142242	[269.14, 273.29]	-0.104820549
63	51	Alabama	<	47	3	1	269.1942806	[266.50, 271.89]	269.4574558	[266.31, 272.61]	268.9201118	[265.93, 271.91]	0.537344031
64	52	District of Columbia	<	51	0	0	265.2593657	[263.46, 267.06]	264.1765343	[261.28, 267.08]	266.2977607	[263.75, 268.84]	-2.121226456
65	NOTE: National public is included for reference only and is not included in sorting the jurisdictions. Score differences are calculated based on differences between unrounded average scale scores.												
66	SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2013 Mathematics Assessment.												

Figures 1a and 1b above: Headers and footers to be deleted.
Confidence interval column with brackets to remove and values to split into two column.

My first editing step is enter column names above the first data row above before I remove the header. Most student are familiar with Excel. In any case on my system with a mouse I click on the first data row, row 11, then right mouse and select insert, the enter the column labels. Figure 2 shows the empty inserted row that has become row 11.

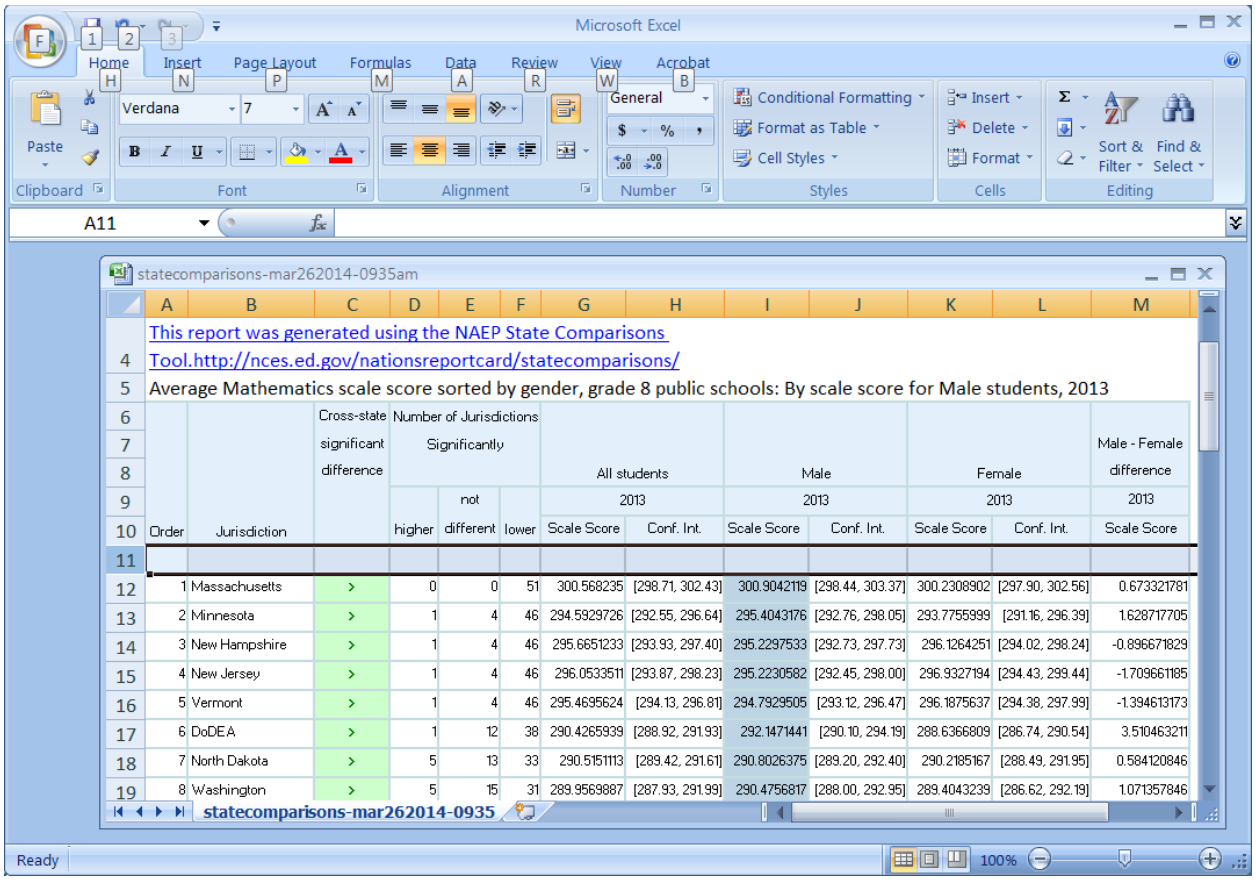
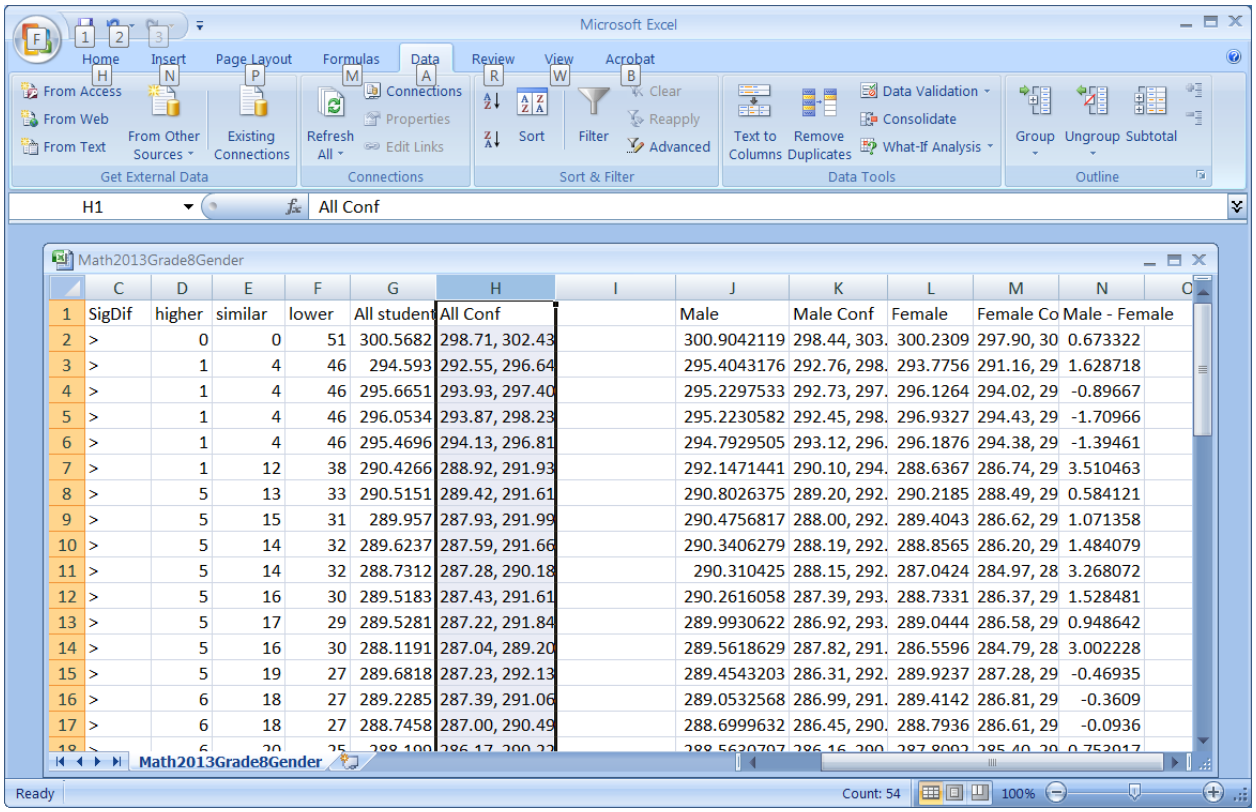


Figure 2. Inserting a row to put column labels before deleting the top 10 rows.

Then select the header rows and delete them and also delete the two footer rows.

The new step is to remove the "[" in the three confidence interval columns. This can be quickly done by selecting a column and using the replace option to replace [with nothing. The top right of Figure 2 shows the Find & Select Menu at the provide access to replace option. After doing this for the three confidence interval column use the replace option to replace] with nothing.

Before splitting the three confidence interval column enter and empty column to right of each confidence interval column. This is where splitting will put the upper confidence bound. To insert an empty column click on the column to right of the confidence interval column and right mouse to chose the insert option.



	C	D	E	F	G	H	I	J	K	L	M	N	O
1	SigDif	higher	similar	lower	All student	All Conf		Male	Male Conf	Female	Female Co	Male - Female	
2	>	0	0	51	300.5682	298.71, 302.43		300.9042119	298.44, 303.	300.2309	297.90, 30	0.673322	
3	>	1	4	46	294.593	292.55, 296.64		295.4043176	292.76, 298.	293.7756	291.16, 29	1.628718	
4	>	1	4	46	295.6651	293.93, 297.40		295.2297533	292.73, 297.	296.1264	294.02, 29	-0.89667	
5	>	1	4	46	296.0534	293.87, 298.23		295.2230582	292.45, 298.	296.9327	294.43, 29	-1.70966	
6	>	1	4	46	295.4696	294.13, 296.81		294.7929505	293.12, 296.	296.1876	294.38, 29	-1.39461	
7	>	1	12	38	290.4266	288.92, 291.93		292.1471441	290.10, 294.	288.6367	286.74, 29	3.510463	
8	>	5	13	33	290.5151	289.42, 291.61		290.8026375	289.20, 292.	290.2185	288.49, 29	0.584121	
9	>	5	15	31	289.957	287.93, 291.99		290.4756817	288.00, 292.	289.4043	286.62, 29	1.071358	
10	>	5	14	32	289.6237	287.59, 291.66		290.3406279	288.19, 292.	288.8565	286.20, 29	1.484079	
11	>	5	14	32	288.7312	287.28, 290.18		290.310425	288.15, 292.	287.0424	284.97, 28	3.268072	
12	>	5	16	30	289.5183	287.43, 291.61		290.2616058	287.39, 293.	288.7331	286.37, 29	1.528481	
13	>	5	17	29	289.5281	287.22, 291.84		289.9930622	286.92, 293.	289.0444	286.58, 29	0.948642	
14	>	5	16	30	288.1191	287.04, 289.20		289.5618629	287.82, 291.	286.5596	284.79, 28	3.002228	
15	>	5	19	27	289.6818	287.23, 292.13		289.4543203	286.31, 292.	289.9237	287.28, 29	-0.46935	
16	>	6	18	27	289.2285	287.39, 291.06		289.0532568	286.99, 291.	289.4142	286.81, 29	-0.3609	
17	>	6	18	27	288.7458	287.00, 290.49		288.6999632	286.45, 290.	288.7936	286.61, 29	-0.0936	

Figure 3. Insert empty columns to the right of each confidence interval column before spitting the column. While not shown above I did this for all three columns before splitting any columns.

To split columns select Excel data tab and select the Text to Column menu that appears above the Male column in Figure 3. Figures 4a, 4b, 4c, and 4d below indicate quick sequence of steps to split columns.

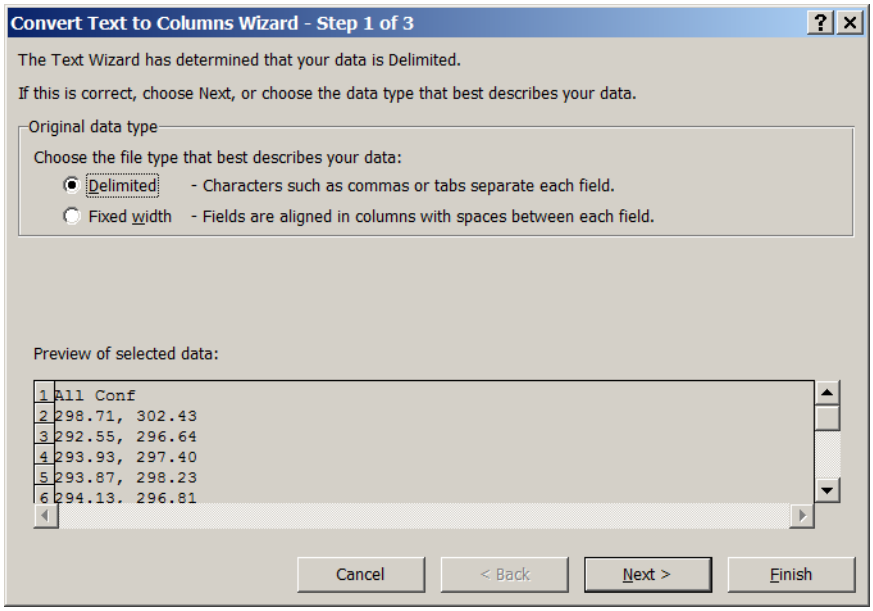


Figure 4a. Click Next

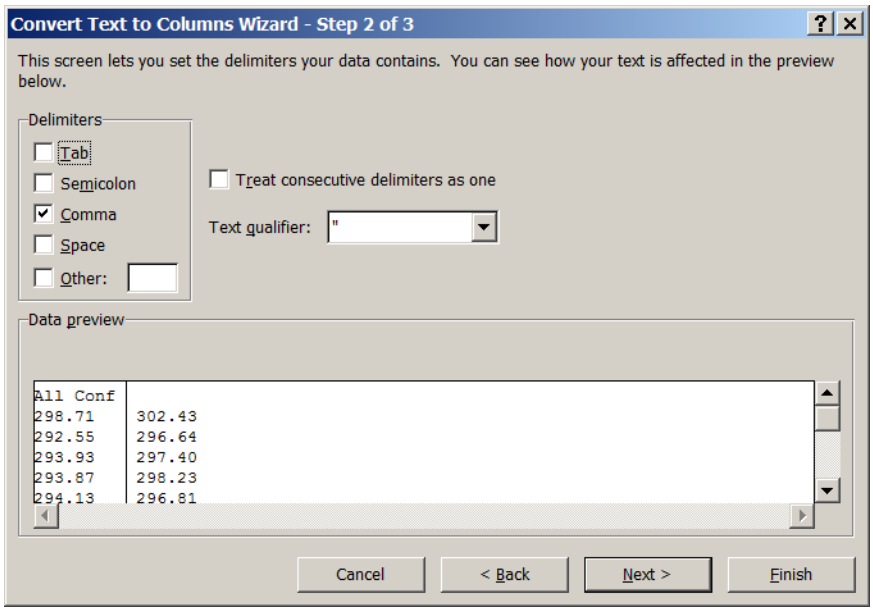


Figure 4b. Change the delimiter from Tab to Comma and click Next

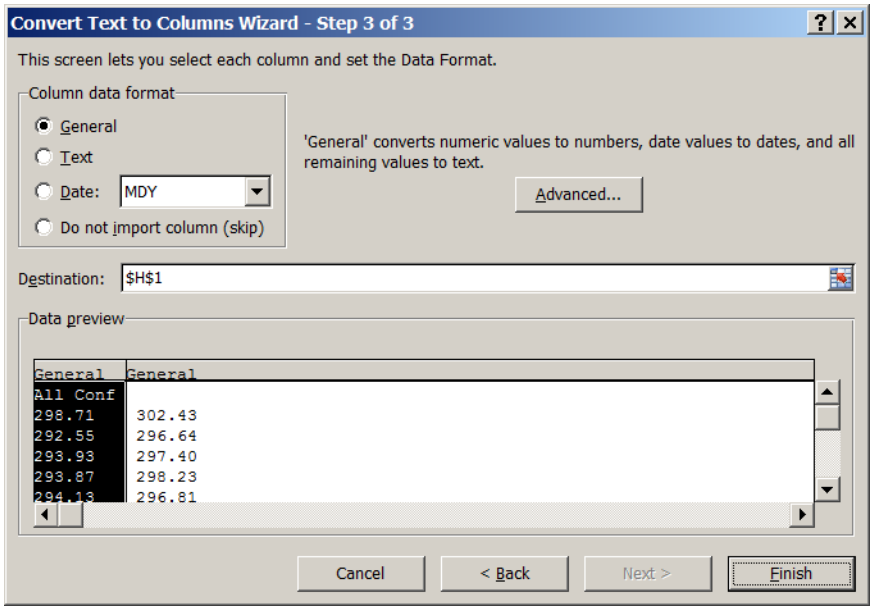


Figure 4c. Click Finish

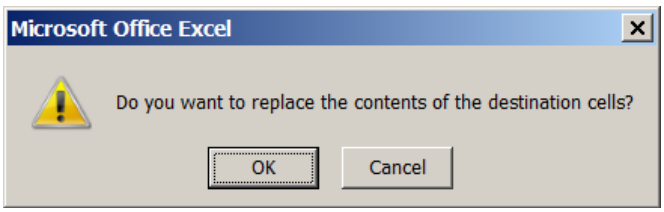
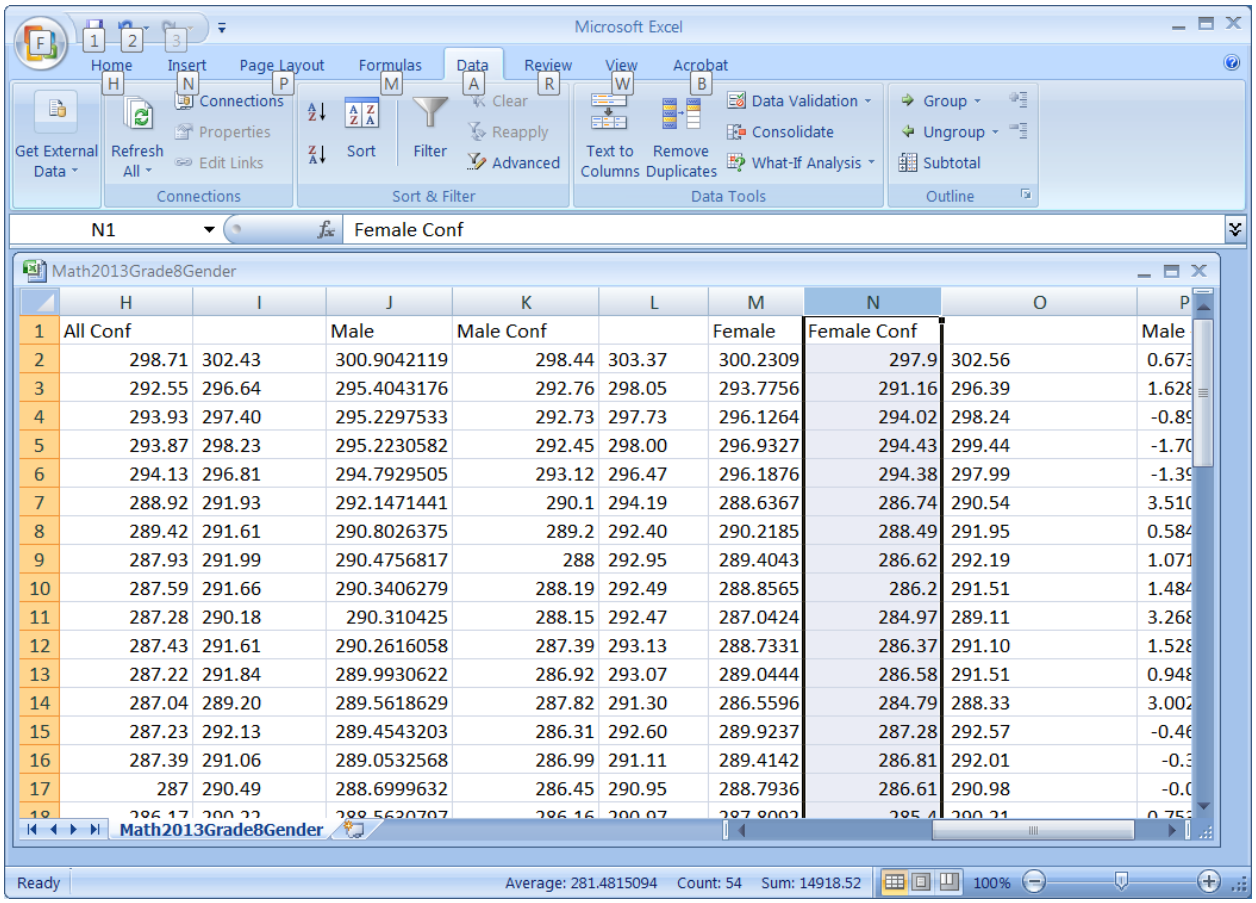


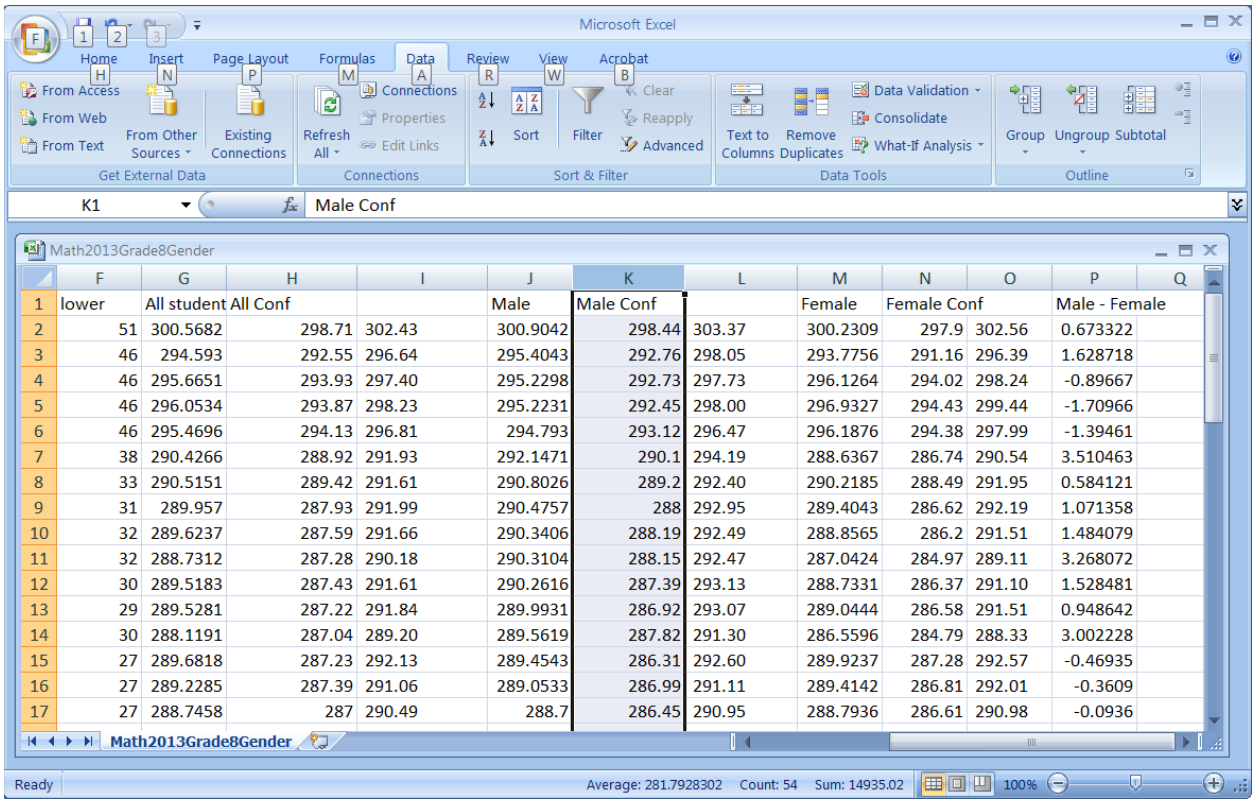
Figure 4d. This may not appear when the right column empty.
I might try cancelling this if it appears.



	H	I	J	K	L	M	N	O	P
1	All Conf		Male	Male Conf		Female	Female Conf		Male
2	298.71	302.43	300.9042119	298.44	303.37	300.2309	297.9	302.56	0.673
3	292.55	296.64	295.4043176	292.76	298.05	293.7756	291.16	296.39	1.628
4	293.93	297.40	295.2297533	292.73	297.73	296.1264	294.02	298.24	-0.89
5	293.87	298.23	295.2230582	292.45	298.00	296.9327	294.43	299.44	-1.70
6	294.13	296.81	294.7929505	293.12	296.47	296.1876	294.38	297.99	-1.39
7	288.92	291.93	292.1471441	290.1	294.19	288.6367	286.74	290.54	3.510
8	289.42	291.61	290.8026375	289.2	292.40	290.2185	288.49	291.95	0.584
9	287.93	291.99	290.4756817	288	292.95	289.4043	286.62	292.19	1.071
10	287.59	291.66	290.3406279	288.19	292.49	288.8565	286.2	291.51	1.484
11	287.28	290.18	290.310425	288.15	292.47	287.0424	284.97	289.11	3.268
12	287.43	291.61	290.2616058	287.39	293.13	288.7331	286.37	291.10	1.528
13	287.22	291.84	289.9930622	286.92	293.07	289.0444	286.58	291.51	0.948
14	287.04	289.20	289.5618629	287.82	291.30	286.5596	284.79	288.33	3.002
15	287.23	292.13	289.4543203	286.31	292.60	289.9237	287.28	292.57	-0.46
16	287.39	291.06	289.0532568	286.99	291.11	289.4142	286.81	292.01	-0.3
17	287	290.49	288.6999632	286.45	290.95	288.7936	286.61	290.98	-0.0

Figure 5. The lacks labels for the three new columns

The last two file preparation steps are add the missing labels (and possible modify other labels) and save the file. I saved it as a comma separated file with a new name. R can immedialy read .csv files. As indicated in the R Graphics CookBook on page 4 the installing the packages xlxs and gdata enable R to fear .xlsx and .xls files respectively.



Microsoft Excel

Home Insert Page Layout Formulas Data Review View Acrobat

From Access From Web From Text From Other Sources Existing Connections Refresh All Connections Properties Edit Links Sort Filter Clear Reapply Advanced Text to Columns Remove Duplicates Data Validation Consolidate What-If Analysis Group Ungroup Subtotal

Get External Data | Connections | Sort & Filter | Data Tools | Outline

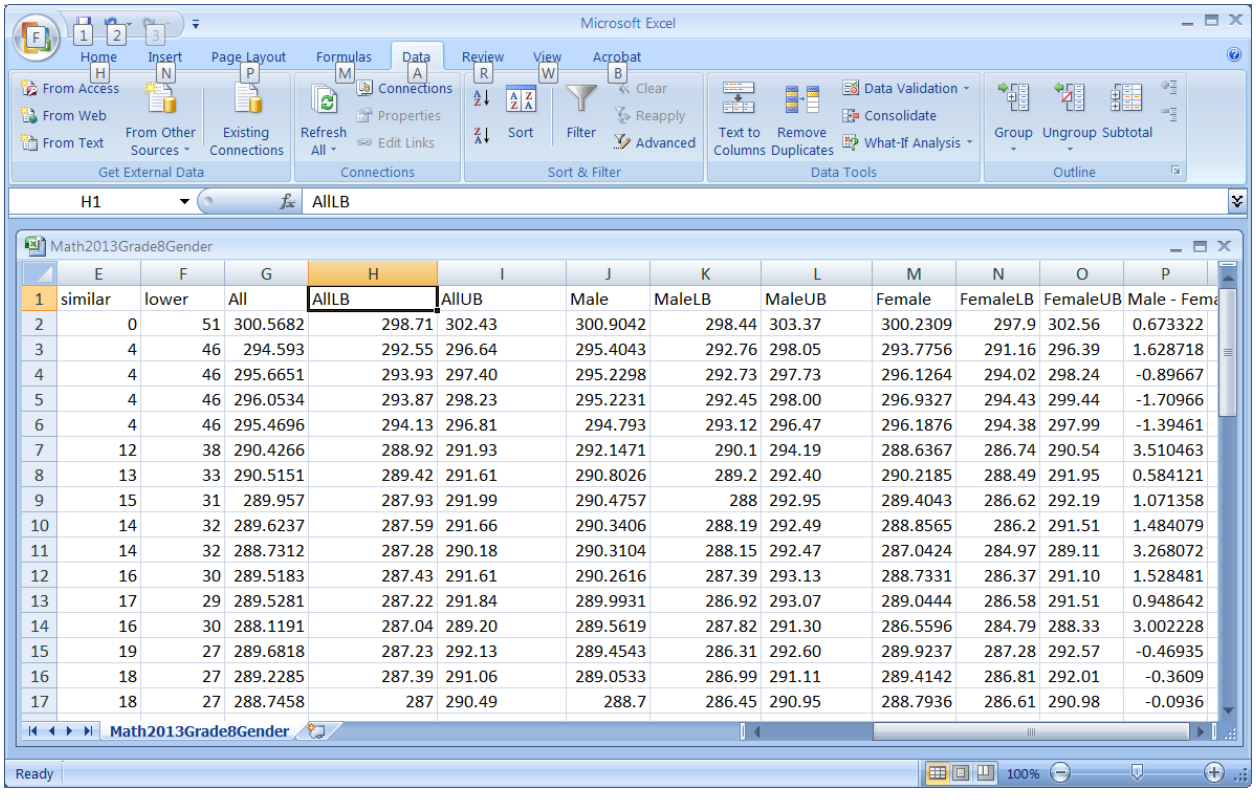
K1 | Male Conf

Math2013Grade8Gender

	F	G	H	I	J	K	L	M	N	O	P	Q
1	lower	All student	All Conf		Male	Male Conf		Female	Female Conf		Male - Female	
2	51	300.5682	298.71	302.43	300.9042	298.44	303.37	300.2309	297.9	302.56	0.673322	
3	46	294.593	292.55	296.64	295.4043	292.76	298.05	293.7756	291.16	296.39	1.628718	
4	46	295.6651	293.93	297.40	295.2298	292.73	297.73	296.1264	294.02	298.24	-0.89667	
5	46	296.0534	293.87	298.23	295.2231	292.45	298.00	296.9327	294.43	299.44	-1.70966	
6	46	295.4696	294.13	296.81	294.793	293.12	296.47	296.1876	294.38	297.99	-1.39461	
7	38	290.4266	288.92	291.93	292.1471	290.1	294.19	288.6367	286.74	290.54	3.510463	
8	33	290.5151	289.42	291.61	290.8026	289.2	292.40	290.2185	288.49	291.95	0.584121	
9	31	289.957	287.93	291.99	290.4757	288	292.95	289.4043	286.62	292.19	1.071358	
10	32	289.6237	287.59	291.66	290.3406	288.19	292.49	288.8565	286.2	291.51	1.484079	
11	32	288.7312	287.28	290.18	290.3104	288.15	292.47	287.0424	284.97	289.11	3.268072	
12	30	289.5183	287.43	291.61	290.2616	287.39	293.13	288.7331	286.37	291.10	1.528481	
13	29	289.5281	287.22	291.84	289.9931	286.92	293.07	289.0444	286.58	291.51	0.948642	
14	30	288.1191	287.04	289.20	289.5619	287.82	291.30	286.5596	284.79	288.33	3.002228	
15	27	289.6818	287.23	292.13	289.4543	286.31	292.60	289.9237	287.28	292.57	-0.46935	
16	27	289.2285	287.39	291.06	289.0533	286.99	291.11	289.4142	286.81	292.01	-0.3609	
17	27	288.7458	287	290.49	288.7	286.45	290.95	288.7936	286.61	290.98	-0.0936	

Math2013Grade8Gender

Ready | Average: 281.7928302 Count: 54 Sum: 14935.02 | 100%



Microsoft Excel

Home Insert Page Layout Formulas Data Review View Acrobat

From Access From Web From Text From Other Sources Existing Connections Refresh All Connections Properties Edit Links Sort Filter Clear Reapply Advanced Text to Columns Remove Duplicates Data Validation Consolidate What-If Analysis Group Ungroup Subtotal

Get External Data | Connections | Sort & Filter | Data Tools | Outline

H1 | AllLB

Math2013Grade8Gender

	E	F	G	H	I	J	K	L	M	N	O	P
1	similar	lower	All	AllLB	AllUB	Male	MaleLB	MaleUB	Female	FemaleLB	FemaleUB	Male - Fema
2	0	51	300.5682	298.71	302.43	300.9042	298.44	303.37	300.2309	297.9	302.56	0.673322
3	4	46	294.593	292.55	296.64	295.4043	292.76	298.05	293.7756	291.16	296.39	1.628718
4	4	46	295.6651	293.93	297.40	295.2298	292.73	297.73	296.1264	294.02	298.24	-0.89667
5	4	46	296.0534	293.87	298.23	295.2231	292.45	298.00	296.9327	294.43	299.44	-1.70966
6	4	46	295.4696	294.13	296.81	294.793	293.12	296.47	296.1876	294.38	297.99	-1.39461
7	12	38	290.4266	288.92	291.93	292.1471	290.1	294.19	288.6367	286.74	290.54	3.510463
8	13	33	290.5151	289.42	291.61	290.8026	289.2	292.40	290.2185	288.49	291.95	0.584121
9	15	31	289.957	287.93	291.99	290.4757	288	292.95	289.4043	286.62	292.19	1.071358
10	14	32	289.6237	287.59	291.66	290.3406	288.19	292.49	288.8565	286.2	291.51	1.484079
11	14	32	288.7312	287.28	290.18	290.3104	288.15	292.47	287.0424	284.97	289.11	3.268072
12	16	30	289.5183	287.43	291.61	290.2616	287.39	293.13	288.7331	286.37	291.10	1.528481
13	17	29	289.5281	287.22	291.84	289.9931	286.92	293.07	289.0444	286.58	291.51	0.948642
14	16	30	288.1191	287.04	289.20	289.5619	287.82	291.30	286.5596	284.79	288.33	3.002228
15	19	27	289.6818	287.23	292.13	289.4543	286.31	292.60	289.9237	287.28	292.57	-0.46935
16	18	27	289.2285	287.39	291.06	289.0533	286.99	291.11	289.4142	286.81	292.01	-0.3609
17	18	27	288.7458	287	290.49	288.7	286.45	290.95	288.7936	286.61	290.98	-0.0936

Math2013Grade8Gender

Ready | 100%