# DATA CLEANING IN EXCEL FOR ASSIGNMENT #1

#### Excel functions and operations we use are:

- DATEVALUE
- SORT left to right
- Insert Row
- Delete Column
- Paste / Transpose

### The data

The Excel Workbook file, Sales\_GMR\_Trend91q1-16q4.xls, includes four worksheets:

- **Explan**: Explanatory Notes and Table of Contents
- **AD**: All Dwellings (we're not interested in this sheet)
- NS: Non-Strata (almost always Houses)
- S: Strata-title properties (almost always Apartments)

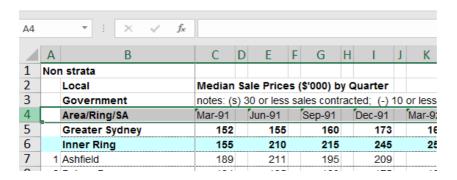
## Things to consider:

- 1. Dates are text, not numeric dates that Excel prefers. (*Note the apostrophe before the text in each date cell.*) We need to change these to proper dates format. (*Excel dates are numbers starting with 1 for Sunday 1, January 1900. Saturday, 18 August 2018 is number 43330*)
- 2. Each data column has an adjoining column for notes. We need to remove those redundant columns.
- 3. Some observations are missing, usually because not enough properties were sold to make a reliable median calculation. We need to be careful that these are not incorrectly interpreted as zeros. (*Missing data is not the same as zero!*)
- 4. Data are in dollar values of the day. They have not been adjusted for inflation. We may want to adjust for inflation.
- 5. There are 51 different LGAs in the data, plus several aggregate region measures. Do we need to consider them all? Probably not.
- 6. Quarterly data are recorded from March 1991 to December 2016. Do we need to go that far back? Probably not.

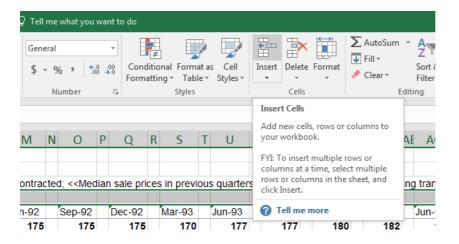
## Transform text date to numeric date

Make a copy of the NS worksheet. Work with the copy.

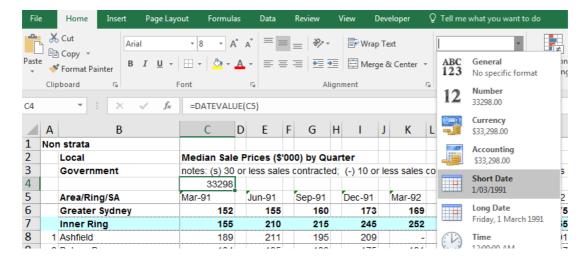
- 1. Insert a new row above the dates.
  - 1) Highlight the whole of *Row#4*. The one with dates. (*Click on the "4" indicating the 4<sup>th</sup> row to highlight the entire row.*)



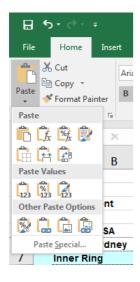
2) Either Right-Mouse-click and select *Insert*, or hit the *Insert* button in the *Ribbon*.



- 3) Select the empty cell above the first date, *Cell D4*.
  - Insert the formula: =DATEVALUE (D5)
  - You should have a number (33298)
- 4) Change the Cell formatting to Date
  - Short Date is quick and easy, but you might want to make a custom format to show only Month and Year.



- 5) Copy the DATEVALUE formula from *Cell D4* across to all cells in the row to *Cell HC4*, for December 2016.
  - Check that the dates match up.
    - i) If not, then you have pasted into the wrong cells. Undo and try again.
- 6) Replace the formulae in Row 4 with Dates.
  - Copy the whole of *Row#4*.
  - In the Ribbon, select Paste / Paste Values...



- Check that the dates still match up.
  - i) If not, then you have pasted into the wrong cell. Undo and try again.
- Check that the formula has been replaced with a date.
  - i) Double-check that the formulae worked in all cases.
    - (1) In my spreadsheet, the last three quarters of 2016 were in proper date format, which produced a #VALUE! error because the DATEVALUE formula was expecting text.
    - (2) In GX4, replace the DATEVALUE formula with simply =GX5

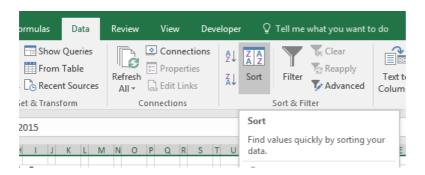
15	#	Dec-15	#	Mar-16	#	#VALUE!	#	#VALUE!	#	#VALUE!	#
5		Dec-15		Mar-16		Jun-16		Sep-16		Dec-16	
90		915		850		885		883		955	
30		1,680	•••••	1,800		1,755		1,825		1,925	
90	ļ	1,580		1,570		1,453		1,640		1,600	
)0	·	1,218		1,255		1,450		1,448		1,450	
25	ļ	1,825		2,030		2,000		2,050		2,350	
25		1,450		1,420		1,500		1,575		1,749	
35		1,220		1,250		1,280		1,319		1,430	
50	· · · · ·	3.060		3 220		3 225		3 825		3 832	

7) Highlight and delete the whole of the redundant dates row, *Row#5*.

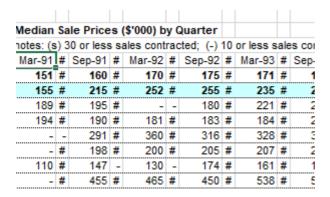
SAVE YOUR WORK.

## Remove redundant additional columns

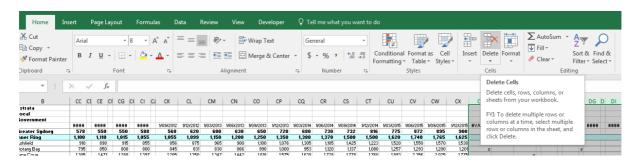
- 1) Highlight the data array that covers all dates and the values for all regions as well as summary values, but not the headings in *Columns A, B, C.* from *Cell D4* to *Cell HC63*
- 2) In the *Ribbon*, hit the *Data* tab
  - a) Select Sort



- 3) Hit the *Options* button,
  - a) select Sort left to right,
  - b) then *OK*.
- 4) In *Sort by*, select *Row 4* (the dates row)
  - a) Then OK.
  - b) Data should all be resorted with dates, in order, on the left and redundant notes columns on the right.
    - i) (You may see columns of hash's. That just means your columns are too narrow to show the numbers inside. Highlight the columns and widen them a little.)
    - ii) Check that the dates are properly sorted to the left, and redundant columns are to the right, with #VALUE! errors all together, probably starting at *Column DA*.



- 5) Hit the *Home* tab in the *Ribbon*.
- 6) Highlight the whole columns relating to the redundant columns, probably from *Column DD* to *Column HC*.
- 7) Hit the *Delete* button in the *Ribbon*.

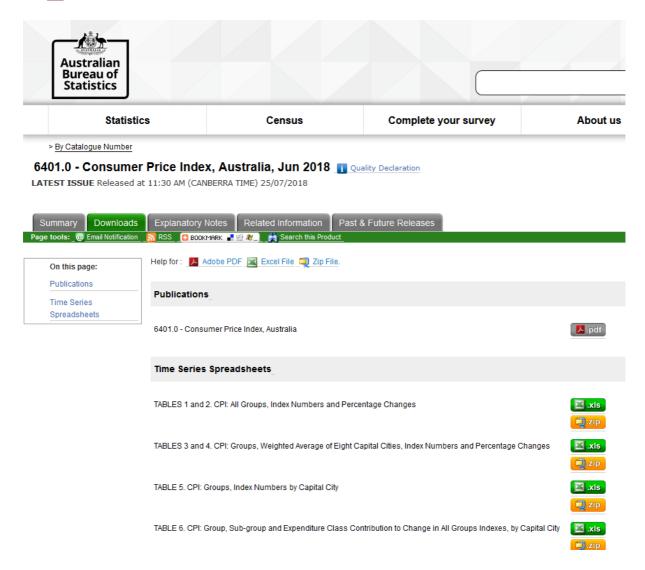


SAVE YOUR WORK.

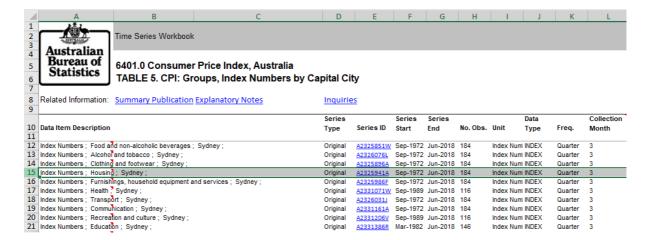
# Adjusting for Inflation

1) Access the latest issue of the Consumer Price Index spreadsheet from the Australian Bureau of Statistics (report 6401)

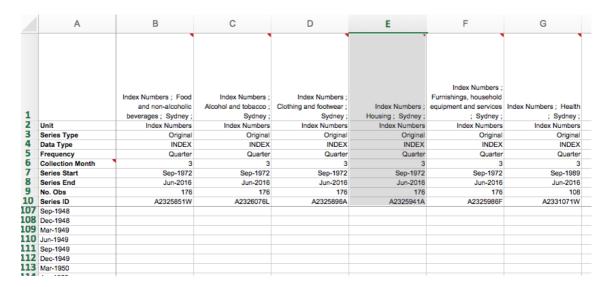
 $\frac{http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/6401.0Main+Features1Jun\%2020}{18}$ 



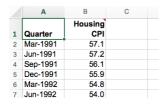
- 2) Download and open the Excel file for *Table 5* (or *Table 9*)
- 3) Choose an appropriate inflation index. I chose *Housing: Sydney*



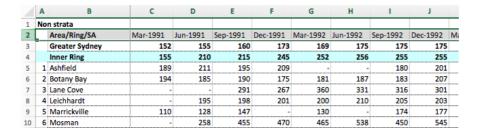
- 4) Hit the hyperlink to take you to the series heading in the *Data1* worksheet.
  - a) Note that this series starts at September 1972, and our Median Housing Prices data starts at March 1991, so there are some redundant rows.



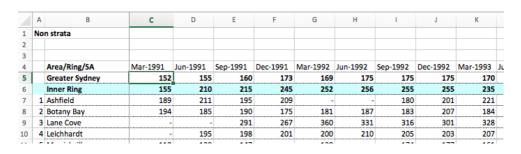
- 5) Copy columns A through to E to a new worksheet
- 6) In the new worksheet,
  - a) delete the Columns B, C, D.
  - b) Delete rows from *Row 2* down to *Row 180* (December 1990), so now we have data that starts at March 1991.
  - c) Relabel the Heading to something shorter and more meaningful.
  - d) This should leave you with two columns, *Date* and *CPI*, that correspond with the Prices data.



7) Return to your *Median Prices* data worksheet.

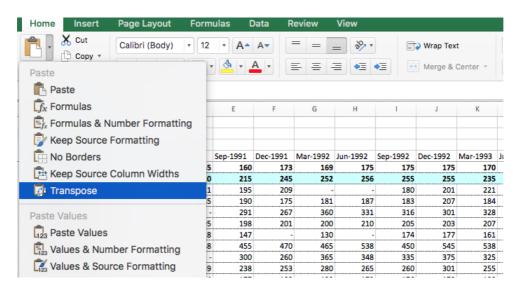


8) Insert two new rows above the current Date row.



#### Copy the CPI columns into rows

- 1) Return to your CPI worksheet
- 2) Copy the two columns for Date and CPI
- 3) Return to your Median Prices worksheet
- 4) Highlight the cell in Column B uppermost of your new rows.
  - a) (Cell B2 in my worksheet)
- 5) In the *Ribbon*, find the *Paste* button, and hit the Down-Arrow on the Paste button to see a drop-down list.



6) Select Transpose

	Α	В	С	D	E	F	G	Н	1	J	K	
1	Non strata											
2		Quarter	Mar-1991	Jun-1991	Sep-1991	Dec-1991	Mar-1992	Jun-1992	Sep-1992	Dec-1992	Mar-1993	Jur
3		Housing CPI	57.1	57.2	56.1	55.9	54.8	54.0	53.3	53.0	53.3	
4		Area/Ring/SA	Mar-1991	Jun-1991	Sep-1991	Dec-1991	Mar-1992	Jun-1992	Sep-1992	Dec-1992	Mar-1993	Jun
5		Greater Sydney	152	155	160	173	169	175	175	175	170	
6		Inner Ring	155	210	215	245	252	256	255	255	235	
7	1	Ashfield	189	211	195	209	-	-	180	201	221	
8	2	Botany Bay	194	185	190	175	181	187	183	207	184	
9	3	Lane Cove		-	291	267	360	331	316	301	328	
10	4	Leichhardt	-	195	198	201	200	210	205	203	207	

#### 7) Check that the dates on both of the dates rows correspond with each other!

a) When you are satisfied that the dates correspond, then you can delete one of the dates rows.

SAVE YOUR WORK.		

#### Inflation adjustment calculation

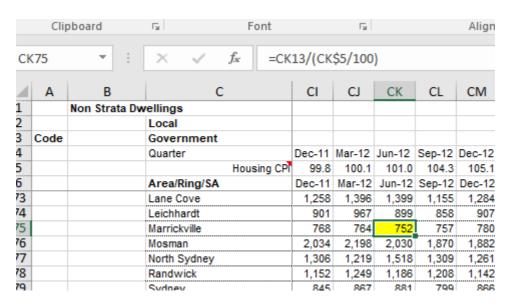
- 1) Make a copy of your Worksheet (just in case)
- 2) Copy the whole data array and paste directly below the original data array.
  - a) This ensures that all headings are in the right places.

54	45	Kiama		150	162	158	170	168	175	145	170	165	160	172	166	184	
55	46	Lake Macquarie	110	115	114	118	120	119	120	120	122	124	121	127	126	127	1
56	47	Maitland	93	95	102	105	108	110	112	108	107	110	109	114	109	108	)
57	48	Newcastle	124	105	110	110	112	110	115	118	115	120	118	120	120	121	
58	49	Port Stephens	114	130	128	135	125	120	110	125	-	-	-	-	-	-	
59	50	Shellharbour	-	114	117	115	124	120	124	125	120	124	123	126	130	129	
50	51	Wollongong		127	127	131										-	
61		Rest of New South Wale	95	98	95	95	97	98	103	102	104	102	103	105	106	108	)
62		NEW SOUTH WALES	112	129	133	132	130	135	135	133	137	140	142	145	145	148	)
53																	
54																	
55		Quarter	Mar-1991	Jun-1991	Sep-1991	Dec-1991	Mar-1992	Jun-1992	Sep-1992	Dec-1992	Mar-1993	Jun-1993	Sep-1993	Dec-1993	Mar-1994	Jun-1994	Se
66		Housing CPI	57.1	57.2	56.1	55.9	54.8	54.0	53.3	53.0	53.3	53.7	53.7	52.8	52.9	53.5	
57		Area/Ring/SA	Mar-1991	Jun-1991	Sep-1991	Dec-1991	Mar-1992	Jun-1992	Sep-1992	Dec-1992	Mar-1993	Jun-1993	Sep-1993	Dec-1993	Mar-1994	Jun-1994	Sep
58		Greater Sydney	152	155	160	173	169	175	175	175	170	177	177	180	182	184	
69		Inner Ring	155	210	215	245	252	256	255	255	235	254	250	270	263	280	J
70	1	Ashfield	189	211	195	209	-	-	180	201	221	215	215	219	226	239	1
71	2	Botany Bay	194	185	190	175	181	187	183	207	184	191	210	225	215	234	
72	3	Lane Cove	-	-	291	267	360	331	316	301	328	338	340	366	366	386	
73	4	Leichhardt	-	195	198	201	200	210	205	203	207	215	217	230	240	249	/

- 3) Highlight the data cell in the upper left of your new data array. (Greater Sydney March 1991)
  - a) In my worksheet, this is Cell D69.
    - i) Yours may be different. That doesn't matter.
  - b) Note, in my worksheet, the first CPI figure is located in *Cell D5*,(57.1 for CPI in March '91) and the first Price datum is in *Cell D7* (for greater Sydney)
    - i) You will need to adjust your formulae accordingly if your first numbers are in different cells.
- 4) In my *Cell D69*, type the formula: =C7/(C\$5/100)
  - a) That is, Adjusted Price equals original price divided by CPI/100. (Dividing by 100 simply changes the Index out of 100 into a decimal.)
  - b) Note the "\$" symbol, which ensures that the row reference is absolute, but column references are relative.

D69 • : × ✓ f <sub>sc</sub> =D7/(D\$5/100)												
4	Α	В	С	D	Е	F	G	Н	1	J	K	
1		Non Strata Dw	vellings									
2			Local	Median 9	Sale Pric	es (\$'00	00) by Q	uarter				
3	Code		Government	notes: (s)	30 or les	ss sales	contract	ed; (-) 1	0 or less	sales c	ontracted	d;
4			Quarter	Mar-91	Jun-91	Sep-91	Dec-91	Mar-92	Jun-92	Sep-92	Dec-92	N
5			Housing CP	57.1	57.2	56.1	55.9	54.8	54.0	53.3	53.0	
6			Area/Ring/SA	Mar-91	Jun-91	Sep-91	Dec-91	Mar-92	Jun-92	Sep-92	Dec-92	N
60	6400	49	Port Stephens	114	130	128	135	125	120	110	125	
61	6900	50	Shellharbour	-	114	117	115	124	120	124	125	
62	8450	51	Wollongong	-	127	127	131	-	-	-	-	L
63	1		GMR	132	143	148	155	152	160	158	155	L
64	1		Rest of New South Wales	95	98	95	95	97	97	103	102	L
65	1		NEW SOUTH WALES	112	129	133	132	130	135	135	133	L
66												
67												
68			Area/Ring/SA	Mar-91			Dec-91	Mar-92	Jun-92	Sep-92	Dec-92	ı
69			Greater Sydney	264				170	175	175	175	
70			Inner Ring	155	210	215	245	252	256	255	255	
71			Ashfield	189	211	195	209	_	_	180	201	L
72			Botany Bay	194	185	190		181	187	183	207	
73			Lane Cove	-	-	291	267	360	331	316	301	

- 5) Check that your formula makes sense.
  - a) \$152 in March 1991 converts to \$266 in January 2012. Yes, that makes sense.
- 6) Copy the formula across to all columns in the data and then down to all rows in the data.
- 7) Check that your conversions make sense.
  - a) Pick a cell further to the right and further down, and check the formula:
    - i) Does it refer to a divisor in *Row 5*?
    - ii) Does it refer to the correct numerator in the corresponding cell for the date and location?
  - b) If the result is not right, then undo, check your formula and your cell references, and try again.

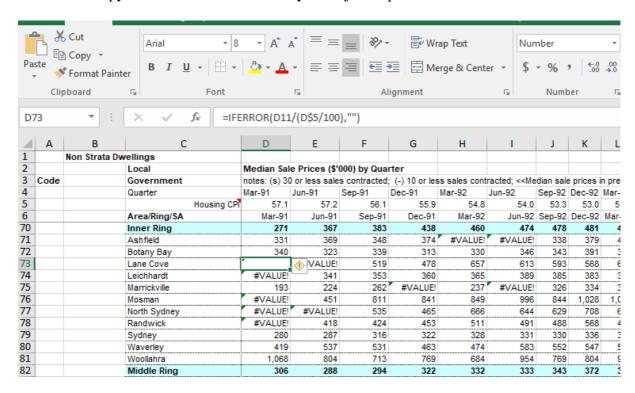


Note that the CPI is arbitrarily set so that January 2012 is 100. So all prices are now in January 2012 prices. With inflation, all prices before January 2012 will go up with adjustment, and prices after January 2012 will go down with adjustment. If you like, you can

change prices to December 2016 prices by multiplying everything by 1.19. Or you could change the January 2012 prices to June 2018 prices by multiplying everything by 1.251. Do you see why? Does it make a difference to your decision-making?

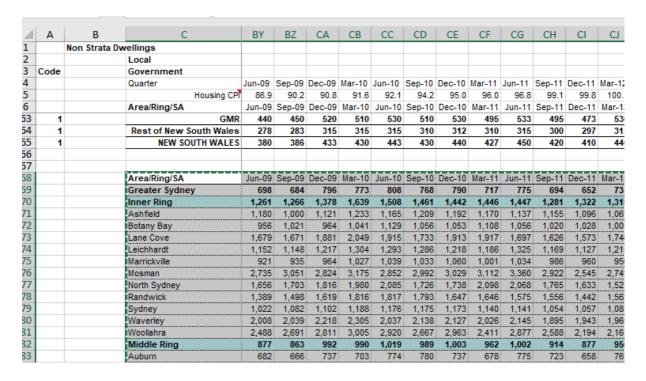
De	59	<b>y</b> : [	× ✓ f <sub>x</sub> =IF	ERROR((D7	//(D\$5/100	))*1.251,""	')
4	Α	В	С	D	E	F	
1		Non Strata Dw	rellings				
2			Local	Median Sa	le Prices (\$'	000) by Qua	rter
3	Code		Government	notes: (s) 3	0 or less sale	s contracted	; (-) 1
4			Quarter	Mar-91	Jun-91	Sep-91	Dec-
5			Housing CP	57.1	57.2	56.1	
6			Area/Ring/SA	Mar-91	Jun-91	Sep-91	1
66							
67							
68			Area/Ring/SA	Mar-91	Jun-91	Sep-91	
69			Greater Sydney	331	339	357	
70			Inner Ring	340	459	479	
71			Ashfield	414	461	435	
70	1			T		·	

- 8) Did you notice that missing values produce #VALUE! errors in your calculations?
  - You can remove these errors by wrapping the CPI adjustment formula in an IFERROR command.
  - Copy this formula to all cells in your adjusted prices matrix.

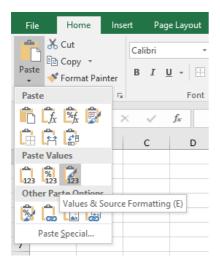


# Save CPI-Adjusted prices to a new worksheet.

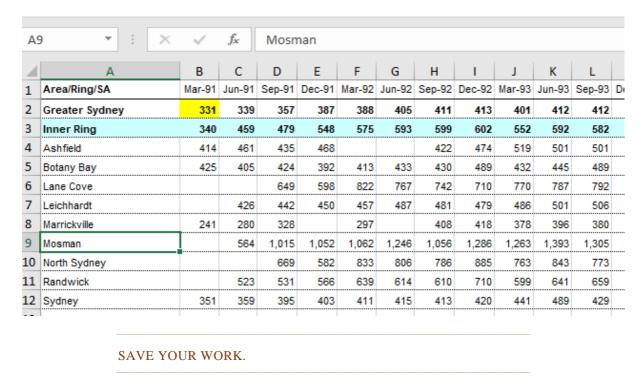
1) Highlight matrix of dates and CPI-adjusted Prices and Region-labels, copy and go to a new worksheet.



2) In Cell A1, hit Paste/Values



- 3) You should have a new dataset of Non-Strata housing prices ready for analysis.
  - Make a copy of this data set and
    - i) remove any rows that you don't need
    - ii) remove any columns that you don't need (think about how far back in time you really need to go for this sort of analysis probably not much)



# Now repeat this exercise with the Strata-title data.

You should be able to copy and paste some rows and some formulae from your Non-Strata data spreadsheet to save time.

SAVE YOUR WORK.	

FEEL PROUD