MIS0855: Data Science
In-Class Exercise for Feb 25-27 – Locating “Bad Data” Using Excel

Objective: Find and fix a data set with incorrect values

Learning Outcomes:

- Use Excel to identify incorrect values and outliers in a data set
- Selectively apply corrections to a data set
- Understand the positive and negative impacts of changing data, even if that change is intended to correct it

In this exercise, you’ll be working with a partial data set of orders for an imaginary company, Vandelay Industries. The data set contains total amount and zip code information for 45,808 orders placed between January, 2009 and January, 2014.

Part 1: Identify incorrect zip codes

1) Download “Vandelay Orders by Zipcode.xlsx” from the class site. Remember where you saved it!

2) Open the file in Microsoft Excel.

3) Take a look through the data (the “Vandelay Order by Zip” tab), and the data dictionary (the “Data Dictionary” tab). The first thing we want to verify is that every zip code in the data set is a valid postal code of the US Postal Service. To do this, we need a list of the correct zip codes. You can find this from various online sources.

   We’ve already imported a list of zip codes into your workbook. You’ll find them under the ZipCodeStateLookup tab. Take a quick look at that tab and check out the data in that sheet.

   You can see it would take a very long time to manually search for each order’s zip code in the lookup table. So we need a quicker way to do that. Do that we will use the MATCH function in Excel.

4) Switch back to the “Vandelay Order By Zip” tab.

5) In cell G1, type “Zip Verify”.

6) In cell G2, type the following formula exactly as it is. Don’t forget the equal sign.

=MATCH(E2,ZipCodeStateLookup!$A$2:$A$42524,0)

then press Enter.

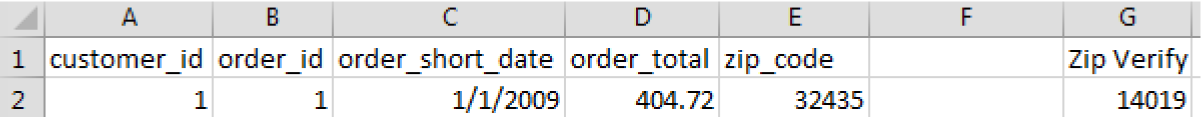
Dissecting the MATCH function (READ THIS – IT’S IMPORTANT!):

MATCH(value, lookup_array, match_type) is an Excel function that searches a list of values for a single value (i.e., looking for the number “105” in a list of house numbers).

So MATCH (E2, ZipCodeStateLookup!$A$2:$A$42524,0) will search for the value in cell E2 (a single zip code) in column A in the ZipCodeStateLookup table (a list of all possible zip codes).

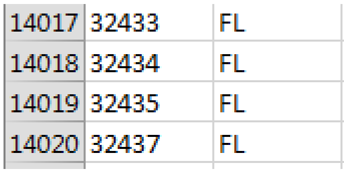
If the value is in the table, it returns the row number where that value is found. If the value isn’t in the table, it returns “#N/A” (an error!). This give us an easy way of checking to see if a value is in a list.

7) In the first row of data, you’ll see this:



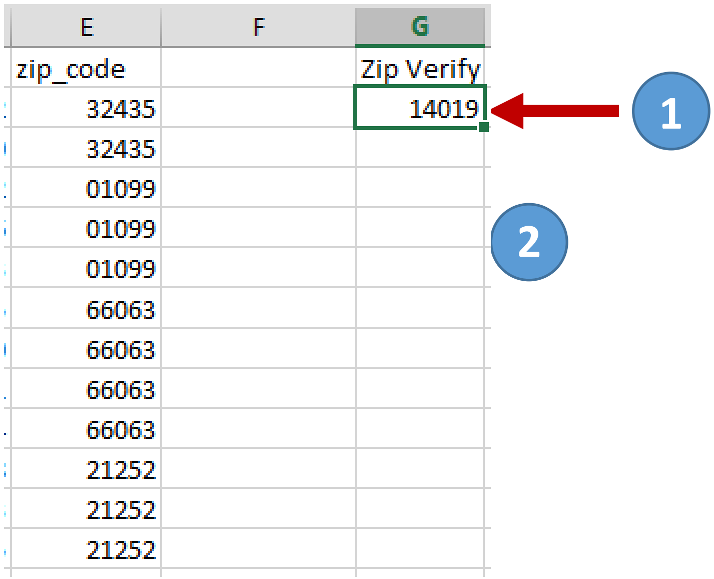
	A	B	C	D	E	F	G
1	customer_id	order_id	order_short_date	order_total	zip_code		Zip Verify
2	1	1	1/1/2009	404.72	32435		14019

8) The ‘14019’ means that it found zip code 32435 in row 14019 of the ZipCodeStateLookup table. To verify that go to the ZipCodeStateLookup tab and scroll down and you’ll find zip code 32435:



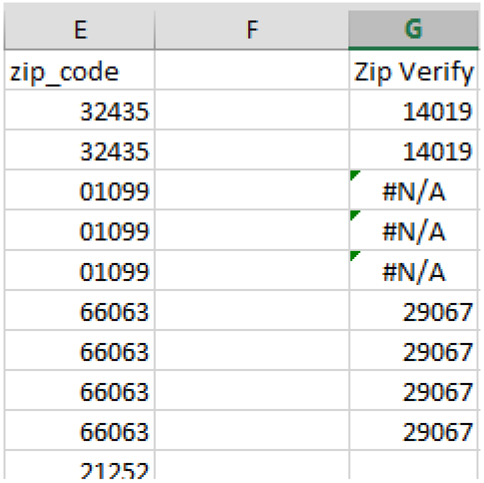
14017	32433	FL
14018	32434	FL
14019	32435	FL
14020	32437	FL

9) Now copy the formula to the next few cells. Click on cell G2 and drag the handle down to cell G10.



E	F	G
zip_code		Zip Verify
32435		14019
32435		
01099		
01099		
01099		
66063		
66063		
66063		
66063		
21252		
21252		
21252		

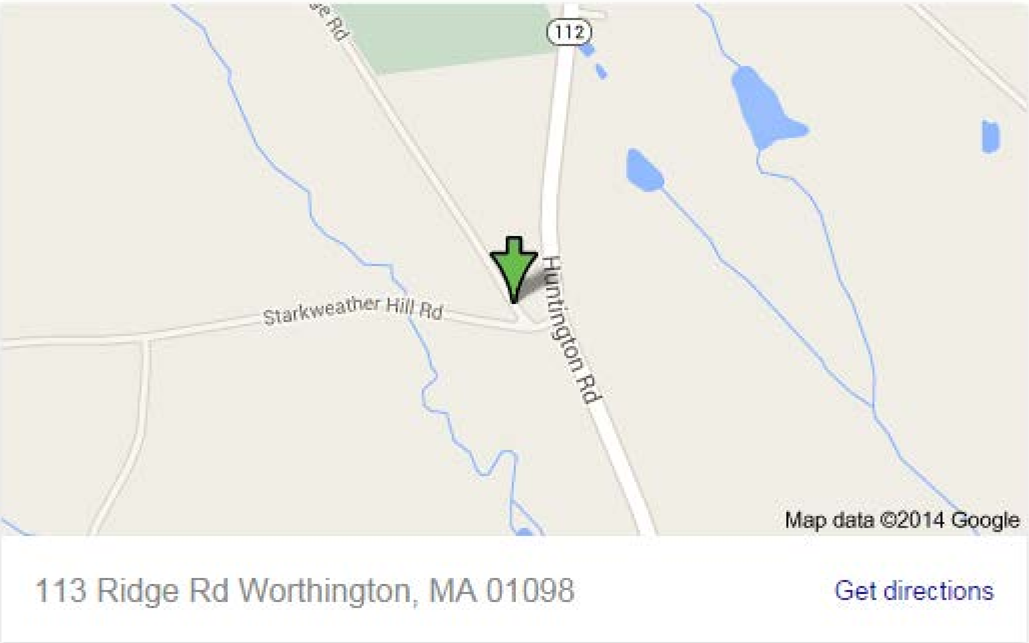
10) You'll see some of the cells have a "#N/A" value. This means that those zip codes weren't found in the official table (there's no row where that value exists) and therefore aren't valid.



E	F	G
zip_code		Zip Verify
32435		14019
32435		14019
01099		#N/A
01099		#N/A
01099		#N/A
66063		29067
66063		29067
66063		29067
66063		29067
21252		

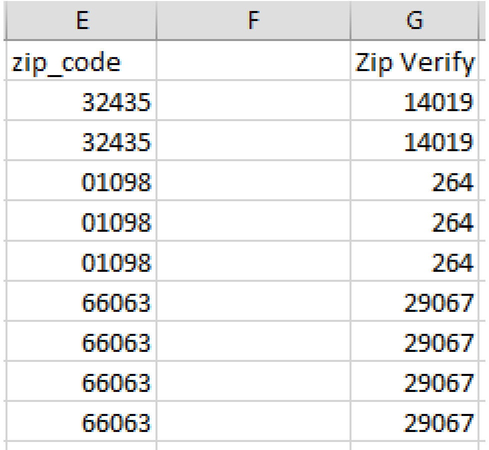
So zip code 01099 is a problem. You would either need to look up the customer (customer_id #2 according to the table) in another database to get the correct zip code, or call the customer to re-verify their address.

11) Assume you've found the address and it's "113 Ridge Road, Worthington, MA." Open a browser and Google that address. You will see this:



113 Ridge Rd Worthington, MA 01098

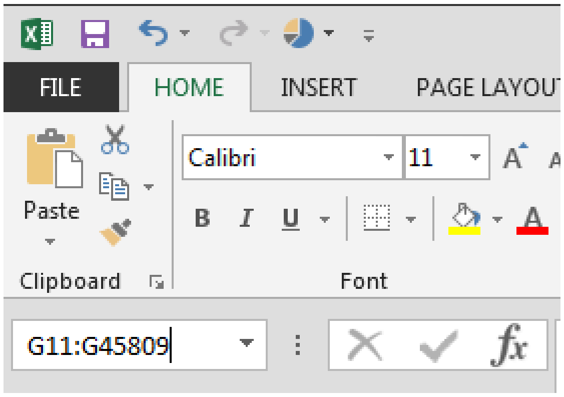
12) Now that you've found the correct zip code (01098), replace cells E3, E4, and E5 with the correct value 01098. The Zip Verify column will now show a number instead of #N/A. This means it's a valid zip code in the list.



E	F	G
zip_code		Zip Verify
32435		14019
32435		14019
01098		264
01098		264
01098		264
66063		29067
66063		29067
66063		29067
66063		29067

13) Now finish copying the values to the rest of the rows. Since there are more than 45,000 rows, dragging is a little impractical. So right-click on cell G10 and select Copy.

14) Now in the name box in Excel type G11:G45809. It should look like this:

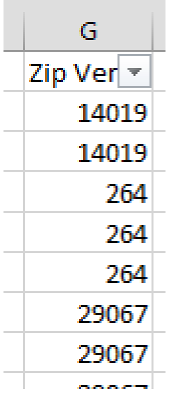


G11:G45809

15) Press Enter and you'll see the range selected. Then right-click and select Paste, and the formula will fill all the way down. Scroll down to verify that's true.

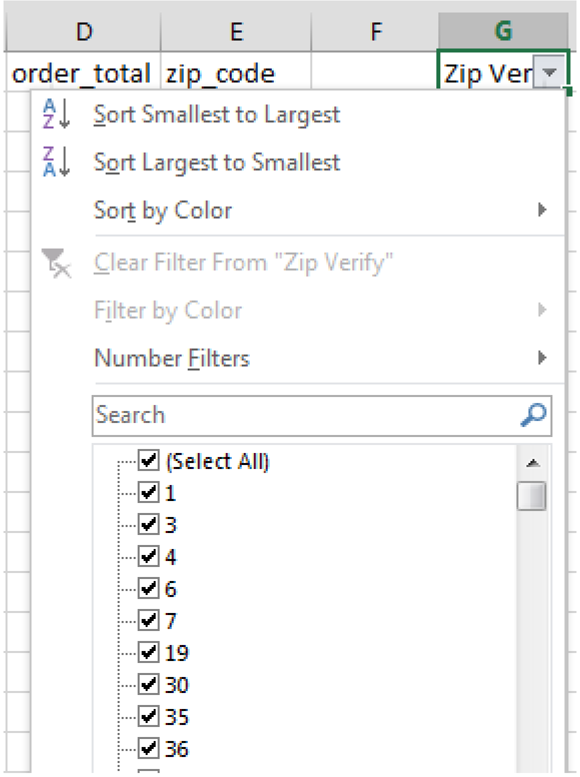
16) To find the other incorrect zip codes, you can now filter the results on the "#N/A" value. Click on cell G1, and then go to the DATA tab (at the top of the window) and then Filter.

You'll see something like this:

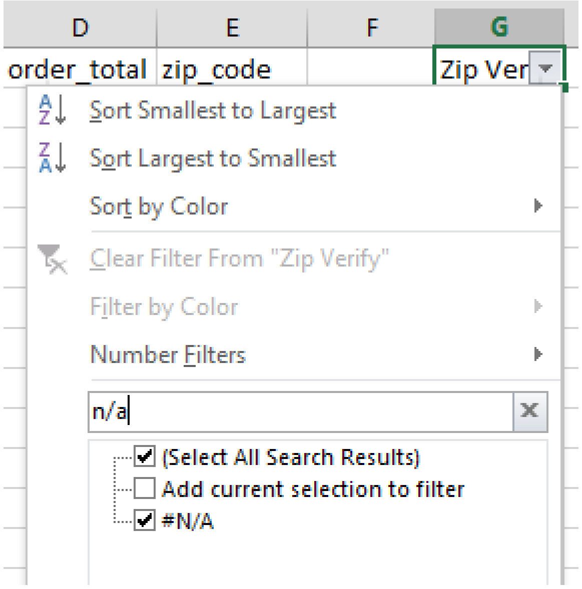


G
Zip Ver
14019
14019
264
264
264
29067
29067

17) Click on the down arrow next to the label, and you’ll see a list of possible values:



18) In the search box, type N/A. You will see this:



19) Click OK and you'll only see the entries with #N/A as a value. Verify that there are 32 rows in the filtered data.

    These are the customers that would have to be double-checked to make sure their Zip Codes were correct.

    Fix the incorrect zip codes using this guide:

Customer	Wrong Zip Code	Right Zip Code
24	42929	42129
204	missing	19087
540	16599	16001
3244	60297	60201
3638	50350	40350
4352	90131	90210
7867	97979	97920
8714	24824	24801

20) Click on the Filter button in the DATA tab to remove the filter.

Part 2: Identify errors in the order total amount

Now we want to figure out if there are any suspicious values for total order amount (order_total). When we talk about suspicious values, we're really talking about outliers – values that are way too low or way too high. In this case, this would include zero dollar order totals (i.e., 0.00) and order totals much larger than the rest.

It's important to identify outliers because they can skew your data because they aren't representative of the rest of the population. They also could be flat-out incorrect; the more atypical the value, the more likely it could be a mistake in the data.

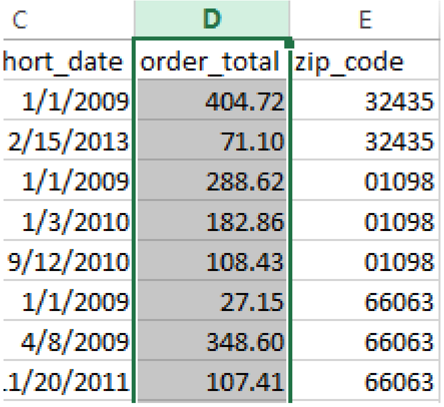
We'll start by looking for order totals much larger than the rest.

1) First, let's determine the current average order total. Click in cell H1 and type:

    =AVERAGE(D:D)

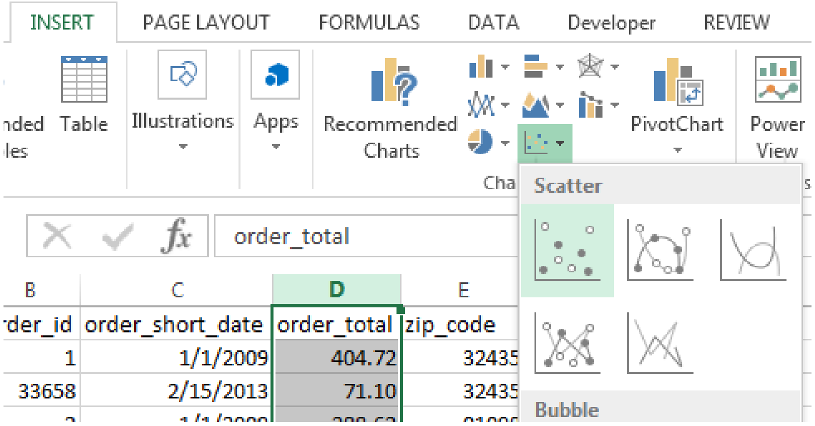
    You'll see the result in the cell: 157.8742. Keep that handy for later.

2) Select Column D (order_total) by clicking on the “D” column header. The entire column should be highlighted:



C	D	E
hort_date	order_total	zip_code
1/1/2009	404.72	32435
2/15/2013	71.10	32435
1/1/2009	288.62	01098
1/3/2010	182.86	01098
9/12/2010	108.43	01098
1/1/2009	27.15	66063
4/8/2009	348.60	66063
1/20/2011	107.41	66063

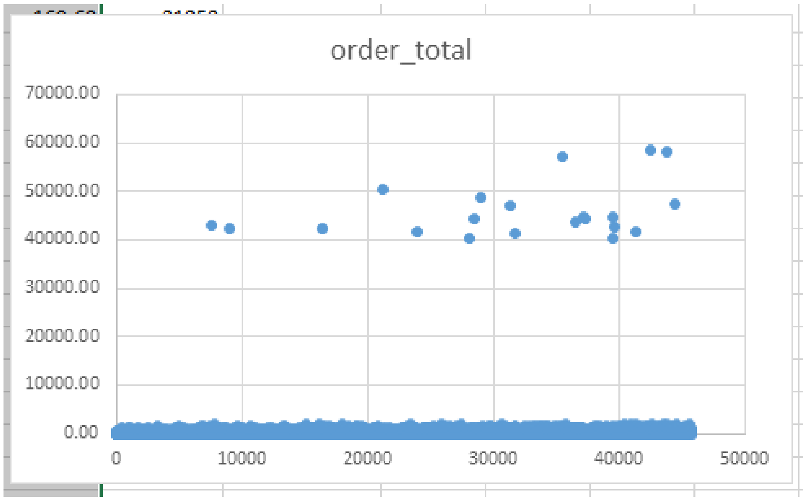
3) Click on the INSERT tab and select the Scatter and Bubble chart icon (⁘ ▾). Select the first Scatter chart (at the top left of the drop-down menu).



order_total

B	C	D	E
der_id	order_short_date	order_total	zip_code
1	1/1/2009	404.72	32435
33658	2/15/2013	71.10	32435

4) Excel will generate and insert a scatter plot into your worksheet.



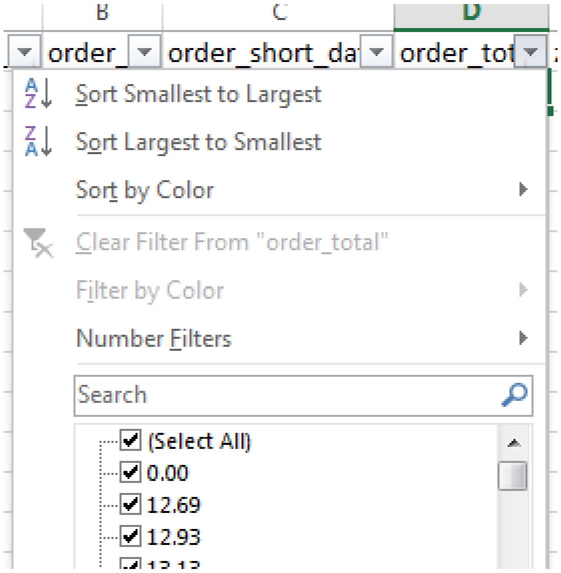
The x-axis (horizontal) doesn't have much meaning – it's just the row number of the data in the spreadsheet. That's why you see the plot end at about 45,000 on the x-axis.

However, the y-axis (vertical) represents the order total. It looks like most order totals are less than about $2,000. However, there is a set of orders that are very large - $40,000 to $60,000. And there is a large gap in-between that group and the rest.

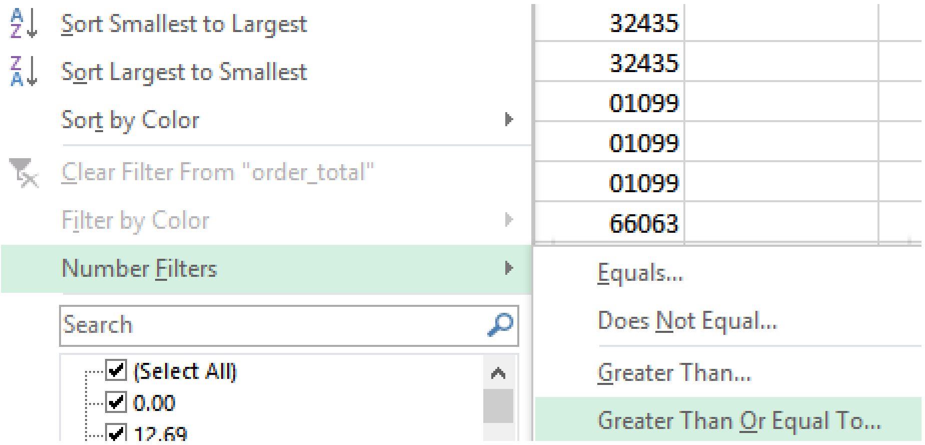
That's suspicious so let's isolate those orders.

5) Click in cell D1.

6) Select the DATA tab and choose Filter. Then select the down arrow in the order_total column. It may take a few seconds to show the menu:

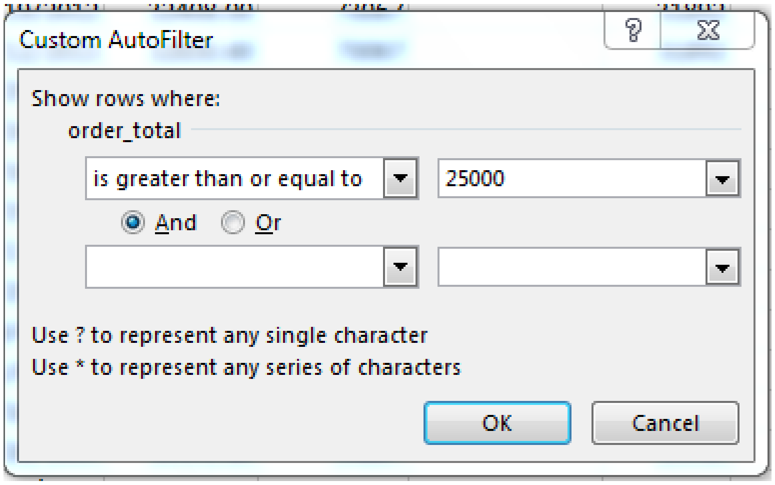


7) Choose "Number Filters" and then "Greater Than Or Equal To…"



8) Fill in “25000” for the value and then click OK. That will be sure to get all of the outliers.

(You could have also chosen “35000,” or “32000,” or “10000.” Any number within that gap in the two sets of data points would work.)



Custom AutoFilter

Show rows where:

order_total

is greater than or equal to ▾ 25000 ▾

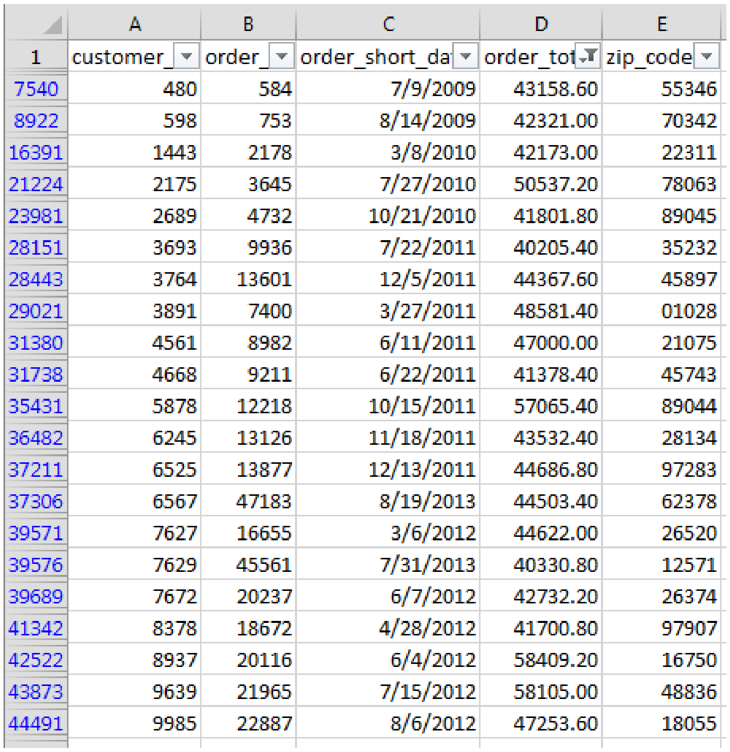
● And ○ Or

▾ ▾

Use ? to represent any single character

Use * to represent any series of characters

9) You’ll see 28 rows left out of the original 45,808. You should also see the average amount still in cell H1 (157.8742).



	A	B	C	D	E
1	customer_	order_	order_short_da	order_tot	zip_code
7540	480	584	7/9/2009	43158.60	55346
8922	598	753	8/14/2009	42321.00	70342
16391	1443	2178	3/8/2010	42173.00	22311
21224	2175	3645	7/27/2010	50537.20	78063
23981	2689	4732	10/21/2010	41801.80	89045
28151	3693	9936	7/22/2011	40205.40	35232
28443	3764	13601	12/5/2011	44367.60	45897
29021	3891	7400	3/27/2011	48581.40	01028
31380	4561	8982	6/11/2011	47000.00	21075
31738	4668	9211	6/22/2011	41378.40	45743
35431	5878	12218	10/15/2011	57065.40	89044
36482	6245	13126	11/18/2011	43532.40	28134
37211	6525	13877	12/13/2011	44686.80	97283
37306	6567	47183	8/19/2013	44503.40	62378
39571	7627	16655	3/6/2012	44622.00	26520
39576	7629	45561	7/31/2013	40330.80	12571
39689	7672	20237	6/7/2012	42732.20	26374
41342	8378	18672	4/28/2012	41700.80	97907
42522	8937	20116	6/4/2012	58409.20	16750
43873	9639	21965	7/15/2012	58105.00	48836
44491	9985	22887	8/6/2012	47253.60	18055

Let’s remove those rows to see how much it affects the average order amount.

10) Highlight the cells in the 21 rows of the order_total data column (D7540 to D44491) and press delete. This will delete the data in those filtered rows, leaving the rest of the data unaffected.

Note that we don't want to get rid of the entire row, just the order_total column. We may want to use the rest of the data, which is correct, for other analyses.

11) You'll see the average amount drop to 136.8824. That's a 15% difference in the overall overage, caused by just 21 data points (about 0.04% of the total sample). This implies that 136.8824 is more representative of the average order price than 157.8742.

A note about deleting those outlier values

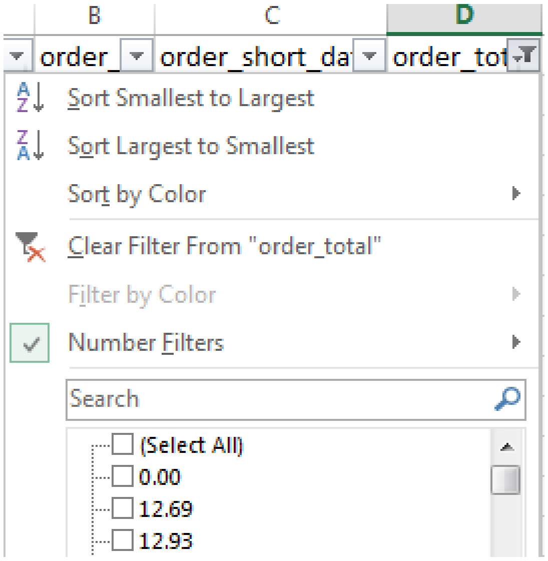
Because we don't know what the right order total values are, we can't correct them. We either have to leave them in, take them out, or make a guess. For right now, we'll leave them out.

However, sometimes that can be bad too. Maybe those order totals aren't wrong: some of our customers may just place really big orders. Or maybe this data is wrong, but it is correctable by looking up the values in the original database of orders placed at the company. By deleting the data you are replacing one potential bias (using incorrect data in the analysis) with another (leaving important data out of the analysis).

What do you do for real? It's a judgment call. Just be ready to explain why you did what you did.

Now let's see if there are $0.00 orders. Those are also likely mistakes in the data, since the company is not supposed to accept empty orders.

12) Click on the Filter for order_total again.



13) Select 0.00 and click OK. We see 213 rows with 0.00 as the order total.

14) Delete those values from the order total column (i.e., D407 to D45578). You'll see the average order amount (cell H1) goes up slightly to 137.5221.

    While we've lost some data – we've gone from 45,808 orders to 45,574 orders – we also now have a much better estimate of the average order.

    NOTE: An alternate way of handling it would be to substitute the average of the rest of the sample (137.5221) for all of those incorrect values (the high values and the zero values). This would allow us to avoid missing data but not alter the overall average.