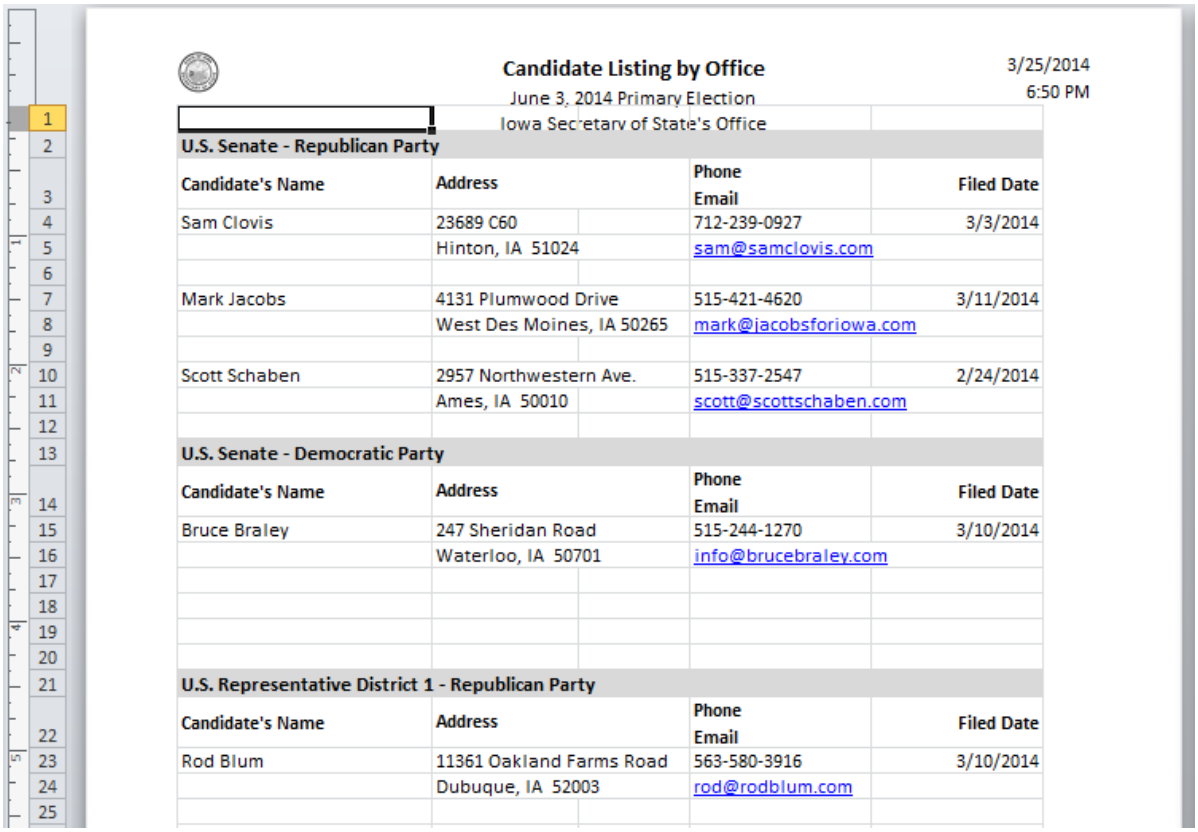
Data Cleaning in Excel: IF statements and more

Jaimi Dowdell, IRE/NICAR

Often, government agencies try to make records look “pretty” by inserting headers, footers and other formatting to line up data into neat columns and rows. While these formatting tricks may be visually pleasing, they don’t always allow us to easily handle data like we would if they were in standard columns and rows. A reporter encountered just this problem when trying to handle a list of candidates running for office in Iowa. Take a look at Figure 1 below or go ahead and open up the file Candidates.xlsx. Figure 2 shows the same data in a much easier to use format.

Figure 1



Candidate Listing by Office

3/25/2014
6:50 PM

June 3, 2014 Primary Election
Iowa Secretary of State's Office

Candidate's Name	Address	Phone / Email	Filed Date
U.S. Senate - Republican Party			
Sam Clovis	23689 C60 Hinton, IA 51024	712-239-0927 sam@samclovis.com	3/3/2014
Mark Jacobs	4131 Plumwood Drive West Des Moines, IA 50265	515-421-4620 mark@jacobsforiowa.com	3/11/2014
Scott Schaben	2957 Northwestern Ave. Ames, IA 50010	515-337-2547 scott@scottschaben.com	2/24/2014
U.S. Senate - Democratic Party			
Bruce Braley	247 Sheridan Road Waterloo, IA 50701	515-244-1270 info@brucebraley.com	3/10/2014
U.S. Representative District 1 - Republican Party			
Rod Blum	11361 Oakland Farms Road Dubuque, IA 52003	563-580-3916 rod@rodblum.com	3/10/2014

Figure 2

	A	B	C	D	E	
1	**Candidate's Name**	**Address**	**PhoneEmail**	**Filed Date**	**Address2**	**Email**
2	Sam Clovis	23689 C60	712-239-0927	41701	Hinton, IA 51024	sam@samclov
3	Mark Jacobs	4131 Plumwood Drive	515-421-4620	41709	West Des Moines, IA 50265	mark@jacobsf
4	Scott Schaben	2957 Northwestern Ave.	515-337-2547	41694	Ames, IA 50010	scott@scottsch
5	Bruce Braley	247 Sheridan Road	515-244-1270	41708	Waterloo, IA 50701	info@brucebra
6	Rod Blum	11361 Oakland Farms Road	563-580-3916	41708	Dubuque, IA 52003	rod@rodblum.
7	Swati Dandekar	2731- 28th Avenue	319-377-2087	41704	Marion, IA 52302	swati@swatid
8	Monica Vernon	326 - 23rd St. Dr. SE	319-431-3970	41705	Cedar Rapids, IA 52403	monicavernon
9	Mariannette Miller-Meeks	11674- 90th St.	641-683-7551	41701	Ottumwa, IA 52501	jane@millerm
10	Joe Grandanette	6215- Gordon Ave	515-710-0798	41705	Des Moines, IA 50312	joegrandanett
11	Staci Appel	10901- 180th Ave	515-238-0033	41709	Ackworth, IA 50001	staci_appel@y
12	Terry E. Branstad	2300 Grand Ave.	515-421-4570	41695	Des Moines, IA 50312	info@branstac
13	Brad Anderson	1525 Beaver Ave.	515-953-9414	41705	Des Moines, IA 50310	brad@andersc

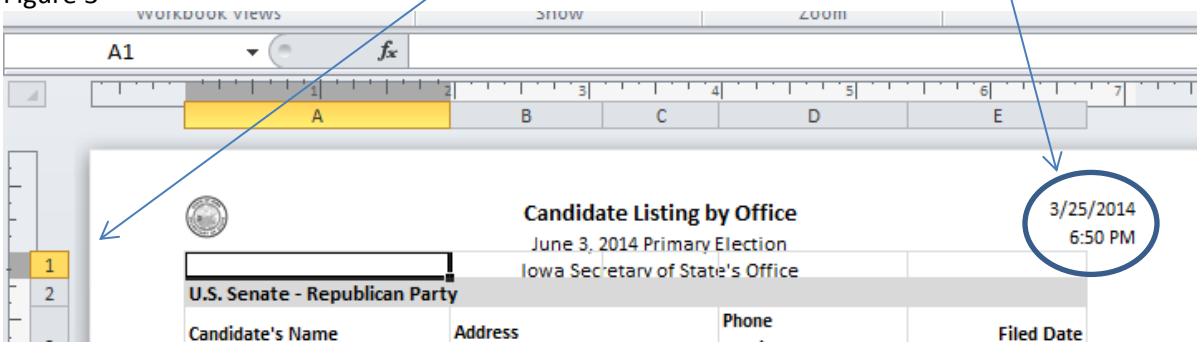
See the difference?

In this exercise, you'll learn how to take the Candidates.xlsx file and make it useable ending up with what you see above. But first, I want you to notice all of the problems with the format originally provided that could hinder things like simple sorts, filters and more in Excel.

Here are two key issues:

1) There is a header "above" where the data rows begin. It includes the date, time and other information that isn't part of the actual data.

Figure 3



2) The information for each candidate isn't on one line. Rather, the address and other contact information are listed on two rows. Also, the office is listed above a group of candidates. Any sorting at all in this current file would separate some of the candidates' information from their record.

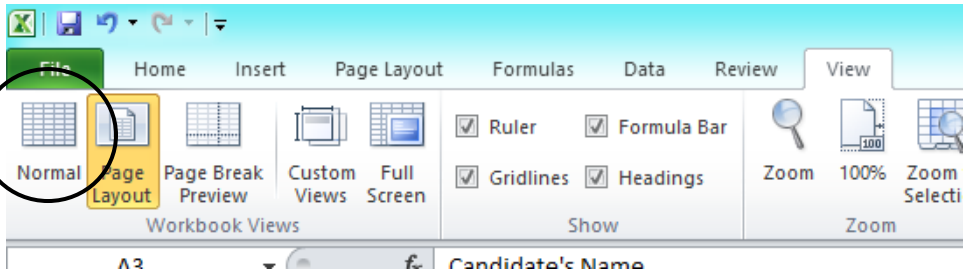
Figure 4

U.S. Senate - Republican Party			
Candidate's Name	**Address**	**Phone Email**	**Filed Date**
Sam Clovis	23689 C60	712-239-0927	3/3/2014
	Hinton, IA 51024	sam@samclovis.com	
Mark Jacobs	4131 Plumwood Drive	515-421-4620	3/11/2014
	West Des Moines, IA 50265	mark@jacobsforiowa.com	
Scott Schaben	2957 Northwestern Ave.	515-337-2547	2/24/2014
	Ames, IA 50010	scott@scottschaben.com	
U.S. Senate - Democratic Party			
		Phone	

Our goal is to get each candidate's information on one line.

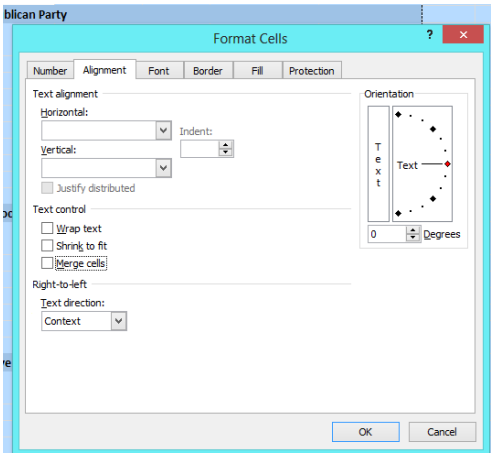
Step 1: Get rid of the header information and make it look like a 'normal' spreadsheet.
Click on the "View" tab in Excel and look to the left-hand side of the page. The file is shown in "Page layout" view. Change that to "normal." That will get rid of the headers and make it look more like a common spreadsheet. See Figure 5.

Figure 5



Step 2: Unmerge cells.
Highlight the entire sheet and right click somewhere on the page. In the menu that appears select "Format cells." Find the "Alignment tab." In the middle of that window you should see the options for "wrap text" and "merge text." They are both filled in. Uncheck both of these as shown in Figure 6 and click OK.

Figure 6



Step 3: Bring the candidate addresses to one line.
Now, we need to bring the information for each candidate up on one line. The first thing we'll move up is the second part of the address. Put your cursor in cell F3 and type in "Address2" to label our new column. Now, look at the pattern of the data. The records always start with the office and party at the top followed by candidate names. There are blanks between where the extra information about each candidate flows down. If you very literally look at it, you could say, "Whenever I see a name in the A column, I know that I need to move other information up a level." For example, I'm going to use this logic in an IF statement to tell Excel that if there is a value in cell A4 that I know the secondary address will need to move up alongside that candidate's record in the F column.

Here's how I write that statement in cell F4:
=IF(A4>0,B5,"")

What this literally says is "If the contents in cell A4 are greater than nothing (i.e., there is a value) then put a copy of whatever is in cell B5 in cell F4. If not, (if there isn't a name in A4), then leave the new "Address2" column blank."

Copy that formula down throughout the document and you should see that the pattern works. You'll notice that we have some additional information "moved up" a line that we don't need. Don't worry; we'll clean that up later.

Figure 7

	A	B	C	D	E	F	G	H
1								
2	**U.S. Senate - Republican Party**							
3	**Candidate's Name**	**Address**		**PhoneEmail**	**Filed Date**	**Address2**		
4	Sam Clovis	23689 C60		712-239-0927	3/3/2014	Hinton, IA 51024		
5		Hinton, IA 51024		sam@samclovis.com				
6								
7	Mark Jacobs	4131 Plumwood Drive		515-421-4620	3/11/2014	West Des Moines, IA 50265		
8		West Des Moines, IA 50265		mark@jacobsforiowa.com				
9								
10	Scott Schaben	2957 Northwestern Ave.		515-337-2547	2/24/2014	Ames, IA 50010		
11		Ames, IA 50010		scott@scottschaben.com				
12								
13	**U.S. Senate - Democratic Party**					Address		
14	**Candidate's Name**	**Address**		**PhoneEmail**	**Filed Date**	247 Sheridan Road		
15	Bruce Braley	247 Sheridan Road		515-244-1270	3/10/2014	Waterloo, IA 50701		
16		Waterloo, IA 50701		info@brucebraley.com				
17								
18								
19								
20								
21	**U.S. Representative District 1 - Republican Party**					Address		
22	**Candidate's Name**	**Address**		**PhoneEmail**	**Filed Date**	11361 Oakland Farms Road		
23	Rod Blum	11361 Oakland Farms Road		563-580-3916	3/10/2014	Dubuque, IA 52003		
24		Dubuque, IA 52003		rod@rodblum.com				
25								

Step 4: Move the email address up alongside the address
Label cell G3 as "email." You'll move each candidate's email next to the address we just moved using the same pattern/method as we used in Step 3.

In cell G4 type this formula:
=IF(A4>0,D5,"")

Notice that it's exactly the same as the previous formula only this time we're selecting the "email" value for the second argument. Copy the formula down throughout the document and you now have each candidate's address and contact information on one line.

Step 5: Cleaning up
Now that you have all of the candidate information on one line, we can clean things up a bit. First, highlight and copy the entire worksheet. Click on the "sheet 2" tab to get a fresh sheet. There, right click in cell A1 and select "paste special" then "values." That'll get rid of the formulas, but you still have them in the first sheet just in case you want to double check your work or if you see a problem later. It also gets rid of unnecessary formatting such as the shading found in the office and party headings.

Step 6: The home stretch

Now that you've pasted the info without the formulas and formatting we can finish it up. Follow these steps to take you (almost) home:

1) Delete the first blank row of the file in sheet 2 so that "U.S. Senate - Republican Party" is now in cell A1. (Remember, to delete a row just right click on the row number at the left and select delete).

2) Select columns A through G and turn on the filter. The reason I want you to highlight the entire columns and not just a chunk of data is that I want to make sure you don't miss out on any records below that you can't see. Look at the filter options for column A. Uncheck the box next to “Select all” to uncheck everything then go all the way to the bottom of the list and select just the blanks. You can do this because you know that if there isn't a candidate name or party listed in column A, we now no longer need that row's information. With the filter on, delete all of the rows that are blank in column A. Next, turn the filter off. That should leave you with 705 rows of information.

3) Turn the filters back on. Look under the column A filter again. This time uncheck everything except for “Candidate’s Name.” This is our header information that is repeated over and over again. Delete the header information for every row except the first one so that at the end you’re left with labels for your columns. We’ll deal with the party and office information in Step 7.

4) But first, notice that column C is completely blank. Double check this with a filter just to make sure there aren’t some records you can’t see. There aren’t, so you can delete that column. This should leave you with data in columns A through F and 434 total rows of data.

Step 7: Add office and party information for each candidate

The last thing you need to do is add the party and office information for each candidate so we make sure we keep the proper details with each candidate. Just like you did before, look at the patterns.

What jumped out at me is the fact that when a party/office is listed columns B, C and D are always blank. See Figure 8 where I’ve highlighted rows with this information for you to better see the patterns.

Figure 8

	A	B	C	D	E	
1	U.S. Senate - Republican Party					
2	**Candidate's Name**	**Address**	**PhoneEmail**	**Filed Date**	**Address2**	**Er**
3	Sam Clovis	23689 C60	712-239-0927	41701	Hinton, IA 51024	sa
4	Mark Jacobs	4131 Plumwood Drive	515-421-4620	41709	West Des Moines, IA 50265	m
5	Scott Schaben	2957 Northwestern Ave.	515-337-2547	41694	Ames, IA 50010	sc
6	U.S. Senate - Democratic Party				Address	Pl
7	Bruce Braley	247 Sheridan Road	515-244-1270	41708	Waterloo, IA 50701	in
8	U.S. Representative District 1 - Republican Party				Address	Pl
9	Rod Blum	11361 Oakland Farms Road	563-580-3916	41708	Dubuque, IA 52003	ro
10	U.S. Representative District 1 - Democratic Party				Address	Pl
11	Swati Dandekar	2731- 28th Avenue	319-377-2087	41704	Marion, IA 52302	sv
12	Monica Vernon	326 - 23rd St. Dr. SE	319-431-3970	41705	Cedar Rapids, IA 52403	m
13	U.S. Representative District 2 - Republican Party				Address	Pl
14	Mariannette Miller-Meeks	11674- 90th St.	641-683-7551	41701	Ottumwa, IA 52501	ja
15	U.S. Representative District 2 - Democratic Party				Address	Pl
16	U.S. Representative District 3 - Republican Party				Address	Pl
17	Joe Grandanette	6215- Gordon Ave	515-710-0798	41705	Des Moines, IA 50312	jo
18	U.S. Representative District 3 - Democratic Party				Address	Pl
19	Staci Appel	10901- 180th Ave	515-238-0033	41709	Ackworth, IA 50001	st
20	U.S. Representative District 4 - Republican Party				Address	Pl

You’ll use this pattern to write one last IF statement to add the data for all of the candidates.

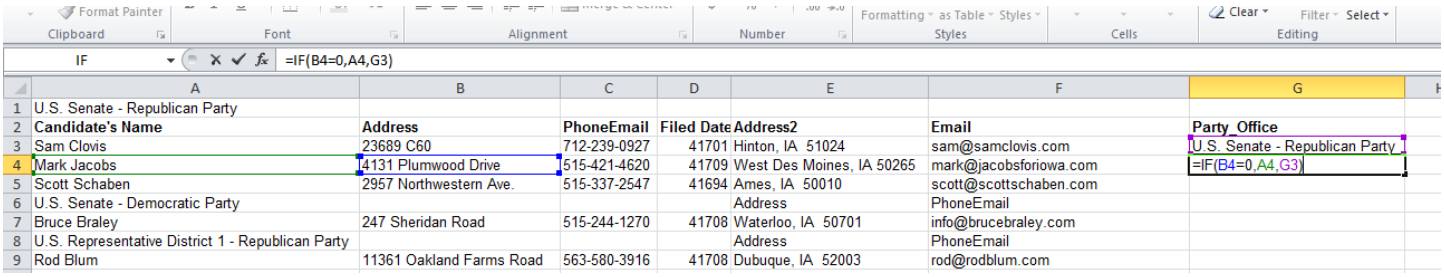
Label cell G2 as something like "Party_Office".

Our first office/party information, "U.S. Senate - Republican Party,” is listed in cell A1. To make things a bit easier, copy that and paste it into cell G3 which is the record for Sam Clovis.

Now is where the patterns come in. We know that for the next record, row 4, that if cell B4 is blank, then A4 is the party information for that candidate and others until we reach another “blank” in column B. If B4 is filled in, we can assume that we’ve moved on to the next office/party and that he candidate's party is the same party as the record above.

We write this statement like this in cell G4:
=IF(B4=0,A4,G3)

Figure 9



	A	B	C	D	E	F	G
1	U.S. Senate - Republican Party						
2	**Candidate's Name**	**Address**	**PhoneEmail**	**Filed Date**	**Address2**	**Email**	**Party_Office**
3	Sam Clovis	23689 C60	712-239-0927	41701	Hinton, IA 51024	sam@samclovis.com	U.S. Senate - Republican Party
4	Mark Jacobs	4131 Plumwood Drive	515-421-4620	41709	West Des Moines, IA 50265	mark@jacobsforiowa.com	=IF(B4=0,A4,G3)
5	Scott Schaben	2957 Northwestern Ave.	515-337-2547	41694	Ames, IA 50010	scott@scottschaben.com	
6	U.S. Senate - Democratic Party				Address	PhoneEmail	
7	Bruce Braley	247 Sheridan Road	515-244-1270	41708	Waterloo, IA 50701	info@brucebraley.com	
8	U.S. Representative District 1 - Republican Party				Address	PhoneEmail	
9	Rod Blum	11361 Oakland Farms Road	563-580-3916	41708	Dubuque, IA 52003	rod@rodblum.com	

Now copy the formula down and double check to make sure the pattern worked throughout.

After you've looked at it all, you can copy and “paste values” “special” one last time. The last thing you’ll want to do is get rid of the rows that just list the party/office. Turn on your filters and under Column B select the blanks. Delete those rows. Delete the extra row at the top.

You’re done! You should end up with 161 candidates listed and one header row.