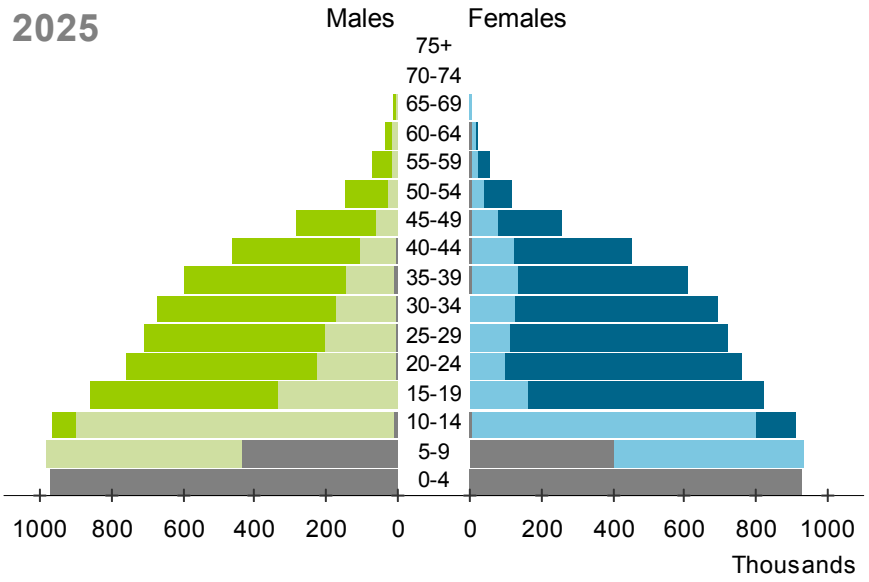
Working with Human Capital Pyramids

When EPDC started producing human capital pyramids, we got the desired effect by using two bar charts that slightly overlap each other. It's a little bit of a pain to get the sizing of the left and right half of the graph to match up, but overall, this method is quite easy. Unfortunately, when using the charts online or in Word documents, it's difficult to get images to overlap correctly. This is why we developed a method to include both halves of the pyramid in a single chart.



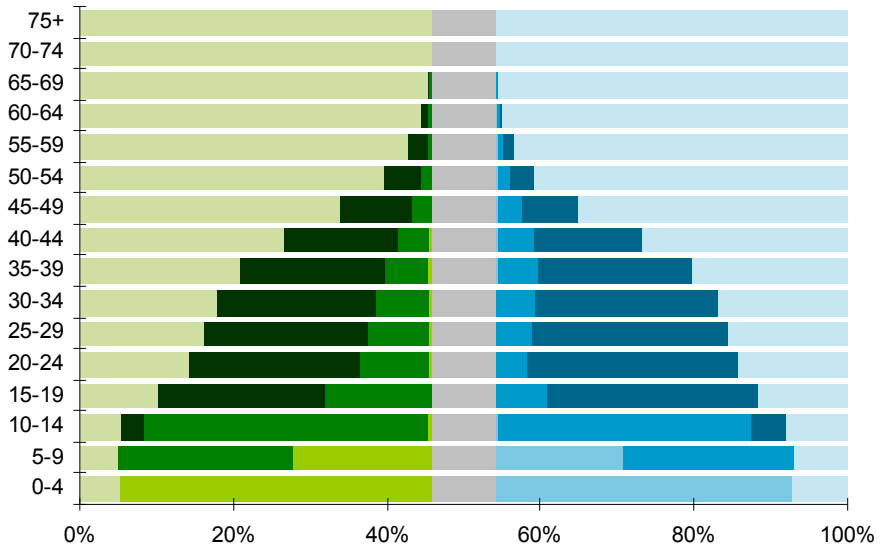
I will go into more detail about the chart, but the basic way that this chart is constructed is with a 100% Stacked Bar chart type combined with an XY Scatter chart. The value axis on the chart is actually an XY series with data labels—we have to fake the value axis in order to have two zeros.

The data is formatted as below. The green columns are the columns with real data. The yellow column with the age groups contains the category labels, and the white columns (*Low*, *Middle*, *High*) are just padding. The *Low*, *Middle*, and *High* series are formatted to be either white or have no fill, so it might not be clear where they are in the pyramid chart above.

		Male			Middle	Female			
AgeGroup	Low	Secondary	Primary	No Schooling		No Schooling	Primary	Secondary	High
0-4	127674	0	0	972326	200000	925305	0	0	174695
5-9	117784	849	547703	433664	200000	399905	529887	2292	167916
10-14	134048	67855	886181	11916	200000	8229	792321	107572	191878
15-19	241571	524673	333756	0	200000	325	162981	655535	281160
20-24	341366	532611	220449	5574	200000	525	98930	657874	342671
25-29	387953	510887	192681	8478	200000	871	112274	609919	376936
30-34	430937	493173	168285	7606	200000	1484	126601	562875	409039
35-39	500759	452858	135237	11146	200000	4392	128775	478115	488717
40-44	634922	358658	102046	4374	200000	4710	117048	333508	644734
45-49	813719	222925	60376	2980	200000	3710	75815	178516	841960
50-54	952184	118851	27388	1576	200000	4104	38110	77917	979869
55-59	1025775	58863	13954	1409	200000	3403	19941	33037	1043620
60-64	1066412	19050	12746	1792	200000	3768	10865	6887	1078479
65-69	1089322	4274	5537	866	200000	923	2771	942	1095365
70-74	1098494	614	773	119	200000	60	154	43	1099743
75+	1099923	28	39	10	200000	1	2	1	1099996

Here is the same pyramid and data, with the series color coded so you can see where they appear in the chart:

AgeGroup	Low	Male			Middle	Female			High
		Secondary	Primary	No Schooling		No Schooling	Primary	Secondary	
0-4	127674	0	0	972326	200000	925305	0	0	174695
5-9	117784	849	547703	433664	200000	399905	529887	2292	167916
10-14	134048	67855	886181	11916	200000	8229	792321	107572	191878
15-19	241571	524673	333756	0	200000	325	162981	655535	281160
20-24	341366	532611	220449	5574	200000	525	98930	657874	342671
25-29	387953	510887	192681	8478	200000	871	112274	609919	376936
30-34	430937	493173	168285	7606	200000	1484	126601	562875	409039
35-39	500759	452858	135237	11146	200000	4392	128775	478115	488717
40-44	634922	358658	102046	4374	200000	4710	117048	333508	644734
45-49	813719	222925	60376	2980	200000	3710	75815	178516	841960
50-54	952184	118851	27388	1576	200000	4104	38110	77917	979869
55-59	1025775	58863	13954	1409	200000	3403	19941	33037	1043620
60-64	1066412	19050	12746	1792	200000	3768	10865	6887	1078479
65-69	1089322	4274	5537	866	200000	923	2771	942	1095365
70-74	1098494	614	773	119	200000	60	154	43	1099743
75+	1099923	28	39	10	200000	1	2	1	1099996



Explanation of some of the calculations

Most of the tricky parts of this graph are related to the scale and the tick mark labels.

Each age category (0-4, 5-9, 10-14, etc.) must have data that adds up to the same number. The values for the primary, secondary, and no schooling series are already determined, so we need to determine how much padding is necessary in the series *Low*, *Middle*, and *High*. Before we can do this, we need to determine the target value that everything should add up to.

We start by taking the maximum population (*No Schooling* + *Primary* + *Secondary*) for males and for females. These values are in cells D2 and H2 in the accompanying Excel sheet. We want the male and female sides of the pyramid to have the same maximum value, so we're only interested in the maximum male or female population value. In the accompanying Excel sheet, the maximum male population is 982,216 and the maximum female population is 932,084, so we're going to base the scale on the male population value (in cell N16 in the Excel file). However, we can't use this as the target value for everything to add up to, because we need to put some padding in the middle of the chart, between the male and female halves, and on the left and

right, so that the population bars don't butt up against the sides of the chart. Also, we would like to have a nice round number for the maximum, so that our tick mark labels will look good.

Cell N17 (highlighted in yellow below) in the Excel sheet determines what number to use for the target value.

dummy axis values	dummy axis labels	dummy axis y values		population numbers rounded
1000000	1000	0		13750
800000	800	0		16500
600000	600	0		19250
400000	400	0		22000
200000	200	0		27500
0	0	0		33000
0	0	0		44000
200000	200	0		55000
400000	400	0		66000
600000	600	0		82500
800000	800	0		96250
1000000	1000	0		110000
	max population	982216		137500
	max population rounded	1100000		165000
	interval value	200000		192500
				220000
				275000
				330000

N17 looks up the smallest number in cells P4:P60 (*population numbers rounded*) which is larger than the maximum population value in N16. I came up with this list of numbers mostly through trial and error, but there is some logic to it. All of the pyramid charts have six number labels along the value axis for male (including 0) and six for female. As long as you keep this setup, then you can come up with new values for P4:P60 with the following rule:

Rounded max value = 5.5*desired interval

In this example, the interval is 200,000 (see the pyramid at top of the document), and 5.5*200,000 is 1,100,000. Following this logic, this is why all of the numbers in P4:P60 seem rather arbitrary. If we chose a nice round number like 1,000,000, then the interval would be 1,000,000/5.5, or 181,810, which does not make for a very nice chart.

Cell N18 (below the yellow cell in the image above) shows the interval; in this case, 200,000.

max female pop		932083.92				dummy axis values	dummy axis labels	dummy axis y values
le	Female					1000000	1000	0
	No Schoolin	Primary	Secondary	High		800000	800	0
000	925305	0	0	=N$17-I6-H6-G6		600000	600	0
000	399905	529887	2292	167916		400000	400	0
000	8229	792321	107572	191878		200000	200	0
000	325	162981	655535	281160		0	0	0
000	525	98930	657874	342671		0	0	0
000	871	112274	609919	376936		200000	200	0
000	1484	126601	562875	409039		400000	400	0
000	4392	128775	478115	488717		600000	600	0
000	4710	117048	333508	644734		800000	800	0
000	3710	75815	178516	841960		1000000	1000	0
000	4104	38110	77917	979869			max population	982216
000	3403	19941	33037	1043620			max population rounded	1100000
000	3768	10865	6887	1078479			interval value	200000
000	923	2771	942	1095365				
000	60	154	43	1099743				
000	1	2	1	1099996				

The *Middle* series used in the chart contains values that are linked to the interval value cell (N18). The *High* and *Low* series are calculated by starting with max population rounded (N17) and subtracting the population values for *No Schooling*, *Primary*, and *Secondary* (see figure above).

If you just use the nine series (*Low*, *Sec*, *Pri*, *No Sch*, *Middle*, *No Sch*, *Pri*, *Sec*, *High*), you will have a pyramid, but no meaningful value axis (because of the chart type that is used, the axis will go from 0 to 100%). We need to fake the axis by adding another series to the chart, on the secondary axis, with the Line chart type. The values for this series are all zeros (*dummy axis y values*). The category names are either *dummy axis values* or *dummy axis labels* in the screenshot above. If you want the units to be thousands, then you should use *dummy axis labels*. If you don't want the labels to be scaled at all, use *dummy axis values*. The real secondary axis is not shown on the chart. The dummy axis series is formatted to look like a real axis, by using category names as data labels, and choosing data markers that look like tick marks.

For a step by step tutorial (although it does not include the dummy value axis), try this: http://peltiertech.com/Excel/Charts/tornadochart.html