IDC IDEA DATA CENTER
Building Capacity for High-Quality *IDEA* Data

Outlier Analyses: Step-by-Step Guide

Authors
Danielle Crain
Chris Lysy

Westat®

January 2018
Version 1.0

IDC *IDEA* DATA CENTER
Building Capacity for High-Quality *IDEA* Data

The *IDEA* Data Center (IDC) created this document under U.S. Department of Education, Office of Special Education Programs Grant No. H373Y130002. Richelle Davis and Meredith Miceli serves as the project officers.

The views expressed herein do not necessarily represent the positions or policies of the U.S. Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service, or enterprise mentioned in this publication is intended or should be inferred. This product is in the public domain. Authorization to reproduce it in whole or in part is granted.

For more information about the *IDEA* Data Center's work and its partners, see www.ideadata.org.

Suggested Citation:

Crain, D., and Lysy, C. (2017, October). *Outlier Analysis: Step-by-Step*. *IDEA* Data Center. Rockville, MD: Westat.

IDEAs that Work
U.S. Office of Special Education Programs

TA&D NETWORK

A Step-by-Step Guide for Completing an Outlier Analysis

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980). Outliers are also referred to as anomalies, abnormalities, deviants, and discordant in the data mining and statistics fields. Often what constitutes a 'sufficient' anomaly is based on subjective judgment. Some anomalies may be embedded in a large amount of "noise" and may not be of interest. Most of the time, the significant deviations are the ones of interest." See *IDEA Data Quality: Outlier Analyses Brief* for more information about the importance of outlier analyses and data quality.

The purpose of this document is to demonstrate three ways of completing outlier analyses. It also demonstrates three ways of displaying the analyses. State staff who want to calculate their outlier analyses using their databases and/or programs can use this guide instead of the *IDEA Data Quality: Outlier Analysis Tool* that provides assistance with automatic calculation of outlier analyses.

There is no single right way to do an outlier analysis. Staff need to choose an approach and be systematic.

The tutorials included in this guide each present different methods staff can use to identify and visualize outliers. Staff should pick the method, or methods, that make the most sense to them. [1]

Identifying an Outlier

Identifying an outlier shat starts with a simple question, "What Is Normal?" The testimonials in this guide can help state staff define what is normal and use data visualization to make identified outliers more noticeable.

A Range of Normal

Outliers are the numbers outside of a range of data considered normal. The following tutorials present different ways of identifying what is normal:

- Tutorial 1: Systematically Determining What Is Normal Using the Interquartile Range
- Tutorial 2: Qualitatively Defining a Normal Range
- Tutorial 3: Simply Sorting

Data Visualization Support for Outlier Analyses

Data visualization can help support outlier analyses by making identified outliers more noticeable. The following tutorials present visual approaches that pair well with any of the first three approaches to calculating outlier analysis.

- Tutorial 4: Heat Maps in Excel
- Tutorial 5: Dot Plots in Excel

[1] Sample Part B data were used to illustrate each tutorial, but these approaches will work with Part C data or any other data source.

- Tutorial 6: Dot Plots in Tableau

A Range of Normal

As stated above, outliers are the numbers outside of a range of data that state staff identified as normal. Here are a few ways to define a normal range.

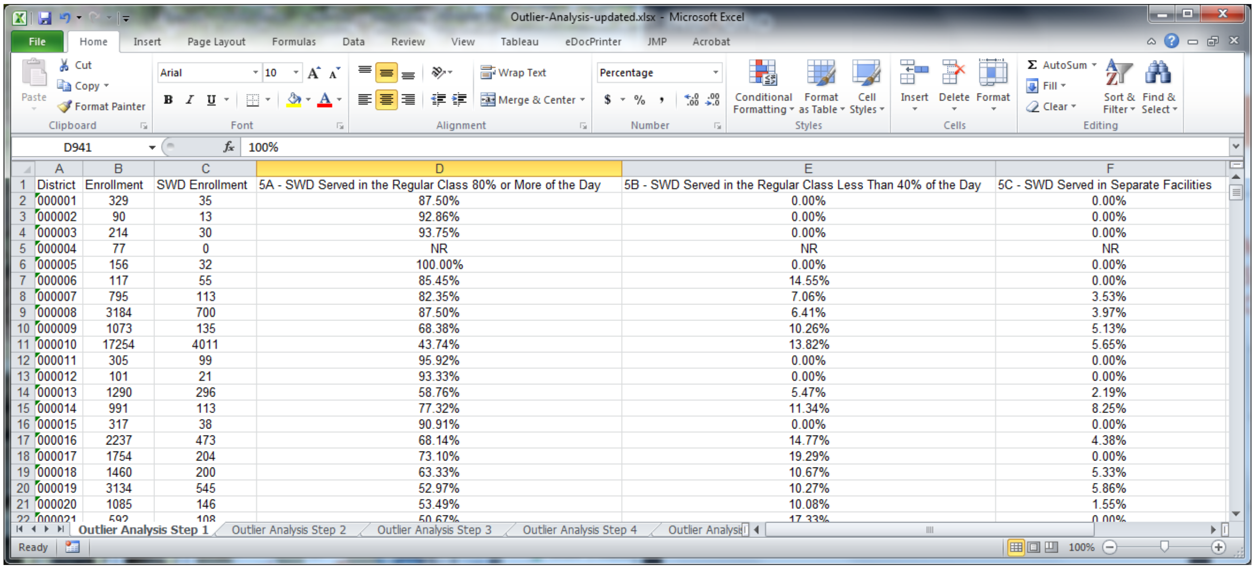
(Note: Within this section, the guide refers to the edges of the normal range as Upper and Lower Fences.)

Tutorial 1: Systematically Determining What Is Normal Using the Interquartile Range

Tutorial 1 is the typical statistical approach to defining what is normal. This approach uses an easy-to-calculate “interquartile range” to identify a normal range for a provided series of data. This series could be data from across all local education agencies (LEAs) and local lead agencies (LLAs) in a state for a single measure. Most likely, state staff will analyze a distribution within a single column in Excel.

Step 1.

To start, staff will need an Excel workbook with at least two columns of data. The following example uses district-level Indicator B5 data.



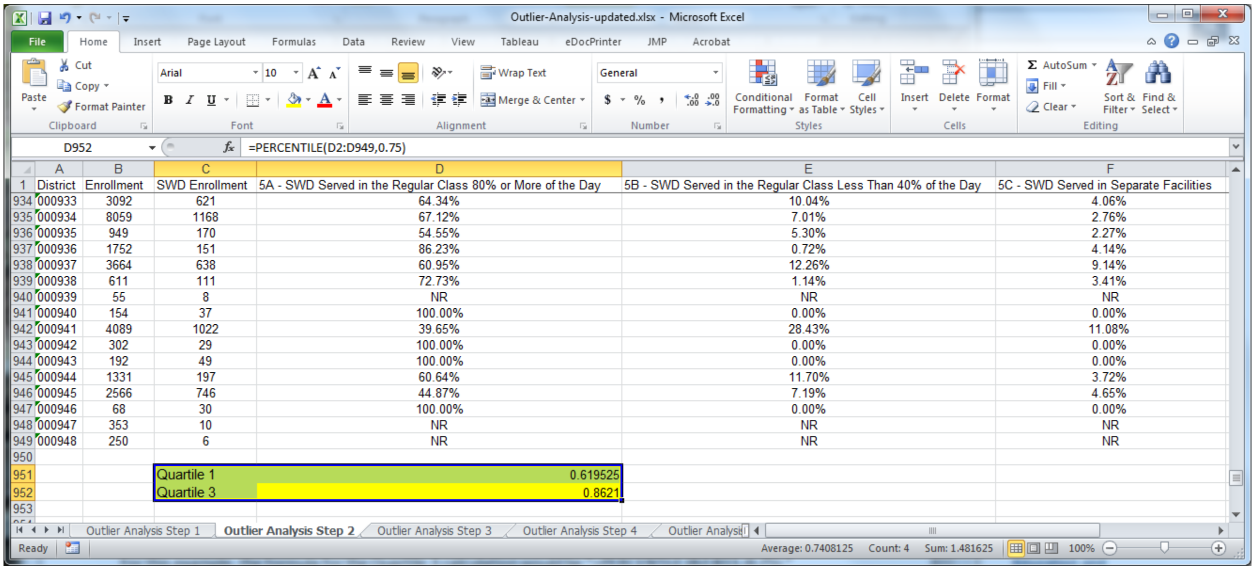
D941 100%

	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilities
2	000001	329	35	87.50%	0.00%	0.00%
3	000002	90	13	92.86%	0.00%	0.00%
4	000003	214	30	93.75%	0.00%	0.00%
5	000004	77	0	NR	NR	NR
6	000005	156	32	100.00%	0.00%	0.00%
7	000006	117	55	85.45%	14.55%	0.00%
8	000007	795	113	82.35%	7.06%	3.53%
9	000008	3184	700	87.50%	6.41%	3.97%
10	000009	1073	135	68.38%	10.26%	5.13%
11	000010	17254	4011	43.74%	13.82%	5.65%
12	000011	305	99	95.92%	0.00%	0.00%
13	000012	101	21	93.33%	0.00%	0.00%
14	000013	1290	296	58.76%	5.47%	2.19%
15	000014	991	113	77.32%	11.34%	8.25%
16	000015	317	38	90.91%	0.00%	0.00%
17	000016	2237	473	68.14%	14.77%	4.38%
18	000017	1754	204	73.10%	19.29%	0.00%
19	000018	1460	200	63.33%	10.67%	5.33%
20	000019	3134	545	52.97%	10.27%	5.86%
21	000020	1085	146	53.49%	10.08%	1.55%
22	000021	592	108	50.67%	17.33%	0.00%

Step 2.

The first calculations are for the first and third quartiles, using the following formulas in Excel: "=percentile(<Cell Range>,0.25)" AND "=percentile(<Cell Range>,0.75)"; staff should just replace <Cell Range> with the range of values, they are checking.

For this example, the formula for the Quartile 1 calculation would be "=PERCENTILE(D2:D949,0.25)," and the Quartile 3 calculation would be "=PERCENTILE(D2:D949,0.75)."



D952 =PERCENTILE(D2:D949,0.75)

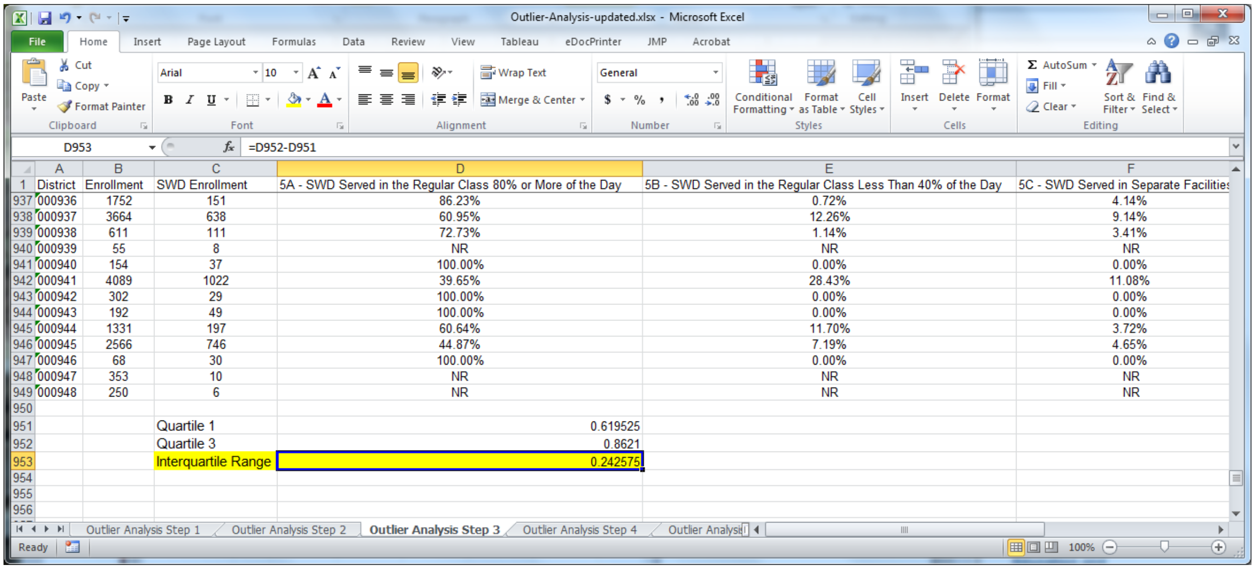
	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilities
934	000933	3092	621	64.34%	10.04%	4.06%
935	000934	8059	1168	67.12%	7.01%	2.76%
936	000935	949	170	54.55%	5.30%	2.27%
937	000936	1752	151	86.23%	0.72%	4.14%
938	000937	3664	638	60.95%	12.26%	9.14%
939	000938	611	111	72.73%	1.14%	3.41%
940	000939	55	8	NR	NR	NR
941	000940	154	37	100.00%	0.00%	0.00%
942	000941	4089	1022	39.65%	28.43%	11.08%
943	000942	302	29	100.00%	0.00%	0.00%
944	000943	192	49	100.00%	0.00%	0.00%
945	000944	1331	197	60.64%	11.70%	3.72%
946	000945	2566	746	44.87%	7.19%	4.65%
947	000946	68	30	100.00%	0.00%	0.00%
948	000947	353	10	NR	NR	NR
949	000948	250	6	NR	NR	NR
950						
951			Quartile 1	0.619525		
952			Quartile 3	0.8621		
953						

Average: 0.7408125 Count: 4 Sum: 1.481625

Step 3.

Next, staff should calculate the interquartile range, which is simply Quartile 3 – Quartile 1.

In this example, the formula would be "=D952-D951."



D953 =D952-D951

	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilities
937	000936	1752	151	86.23%	0.72%	4.14%
938	000937	3664	638	60.95%	12.26%	9.14%
939	000938	611	111	72.73%	1.14%	3.41%
940	000939	55	8	NR	NR	NR
941	000940	154	37	100.00%	0.00%	0.00%
942	000941	4089	1022	39.65%	28.43%	11.08%
943	000942	302	29	100.00%	0.00%	0.00%
944	000943	192	49	100.00%	0.00%	0.00%
945	000944	1331	197	60.64%	11.70%	3.72%
946	000945	2566	746	44.87%	7.19%	4.65%
947	000946	68	30	100.00%	0.00%	0.00%
948	000947	353	10	NR	NR	NR
949	000948	250	6	NR	NR	NR
950						
951			Quartile 1	0.619525		
952			Quartile 3	0.8621		
953			Interquartile Range	0.242575		
954						
955						
956						

Step 4.

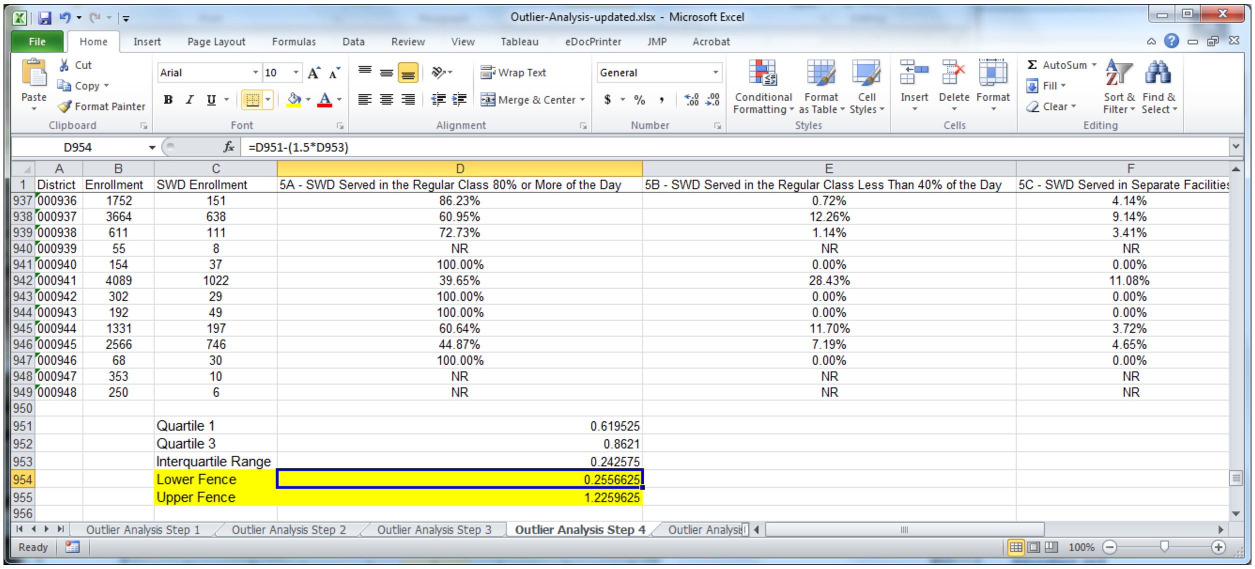
Using the interquartile range (IQR), staff should come up with a “normal range” by setting up fences. The Lower Fence is the bottom of the range, and the Upper Fence is the top.

Calculate the Lower Fence by subtracting 1.5 times the interquartile range from Quartile 1. [Lower Fence = Quartile 1 – (1.5 * IQR)]

Calculate the Upper Fence by adding 1.5 times the interquartile range to Quartile 3. [Upper Fence = Quartile 3 + (1.5 * IQR)]

In this example, the formula for the Lower Fence would be “=D951-(1.5*D953),” and the Upper Fence would be “=D952+(1.5*D953).”

(Note: This is subjective. To create a larger range (fewer outliers), staff should multiply by a larger number than 1.5 (maybe 2 or 3). To create a smaller range, staff should not use a multiplier at all. If staff do not use a multiplier, the definition of the percentile indicates that staff will identify about half of their data points.



D954 =D951-(1.5*D953)

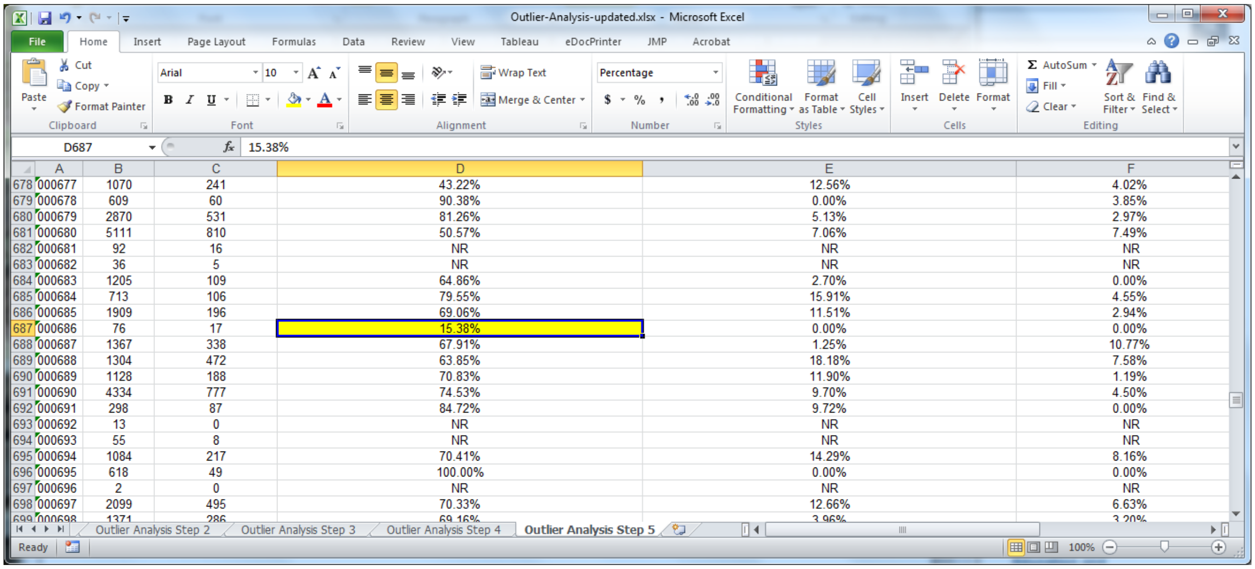
	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilitie
937	000936	1752	151	86.23%	0.72%	4.14%
938	000937	3664	638	60.95%	12.26%	9.14%
939	000938	611	111	72.73%	1.14%	3.41%
940	000939	55	8	NR	NR	NR
941	000940	154	37	100.00%	0.00%	0.00%
942	000941	4089	1022	39.65%	28.43%	11.08%
943	000942	302	29	100.00%	0.00%	0.00%
944	000943	192	49	100.00%	0.00%	0.00%
945	000944	1331	197	60.64%	11.70%	3.72%
946	000945	2566	746	44.87%	7.19%	4.65%
947	000946	68	30	100.00%	0.00%	0.00%
948	000947	353	10	NR	NR	NR
949	000948	250	6	NR	NR	NR
950						
951			Quartile 1	0.619525		
952			Quartile 3	0.8621		
953			Interquartile Range	0.242575		
954			Lower Fence	0.2556625		
955			Upper Fence	1.2259625		
956						

Step 5.

Once the Lower and Upper Fences are known, state staff can identify outliers. An outlier would be any number falling below the Lower Fence or above the Upper Fence.

In this example, that would be any number less than 25.6 percent and any number over 122.6 percent. The value of 15.38 percent would be considered an outlier because it falls below 25.6 percent.

Note: With data that are widely dispersed, it is likely that the Lower Fence could be a negative number or the Upper Fence could fall outside of the possible range of values. Staff could interpret a negative Lower Fence as having no Lower Fence with no low outliers.



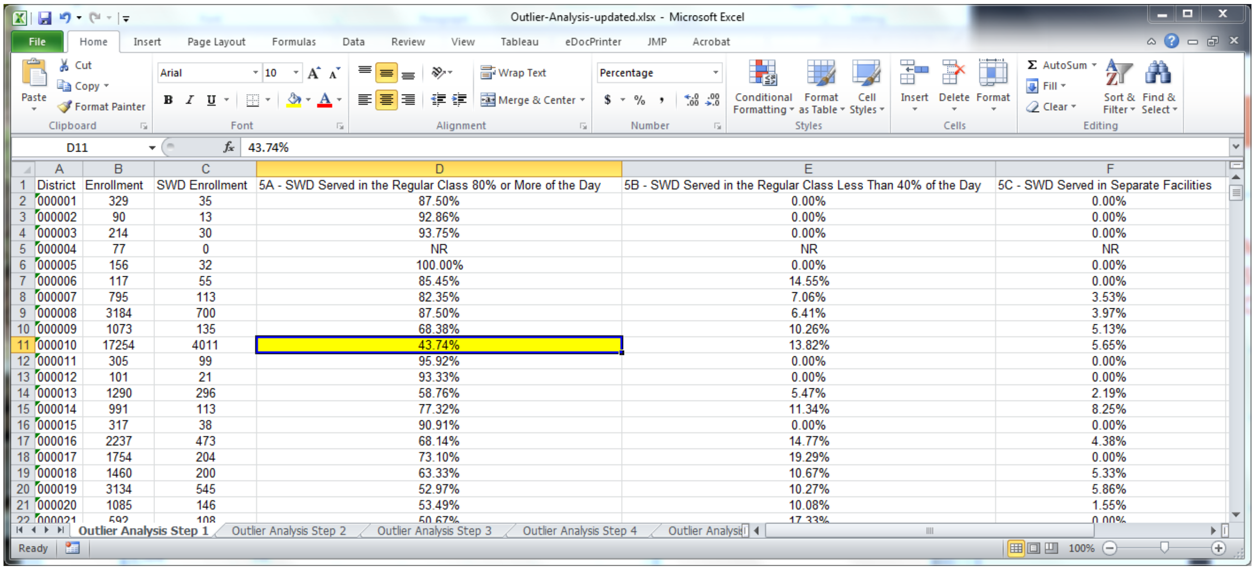
D687 15.38%

	A	B	C	D	E	F
678	000677	1070	241	43.22%	12.56%	4.02%
679	000678	609	60	90.38%	0.00%	3.85%
680	000679	2870	531	81.26%	5.13%	2.97%
681	000680	5111	810	50.57%	7.06%	7.49%
682	000681	92	16	NR	NR	NR
683	000682	36	5	NR	NR	NR
684	000683	1205	109	64.86%	2.70%	0.00%
685	000684	713	106	79.55%	15.91%	4.55%
686	000685	1909	196	69.06%	11.51%	2.94%
687	000686	76	17	15.38%	0.00%	0.00%
688	000687	1367	338	67.91%	1.25%	10.77%
689	000688	1304	472	63.85%	18.18%	7.58%
690	000689	1128	188	70.83%	11.90%	1.19%
691	000690	4334	777	74.53%	9.70%	4.50%
692	000691	298	87	84.72%	9.72%	0.00%
693	000692	13	0	NR	NR	NR
694	000693	55	8	NR	NR	NR
695	000694	1084	217	70.41%	14.29%	8.16%
696	000695	618	49	100.00%	0.00%	0.00%
697	000696	2	0	NR	NR	NR
698	000697	2099	495	70.33%	12.66%	6.63%
699	000698	1371	286	69.16%	3.96%	3.20%

Tutorial 2: Qualitatively Defining a Normal Range

Another way to identify outliers would be to use qualitative information. State staff can identify the numbers that seem out of place given what they know about their state's data, or they can ask colleagues about their own expectations based on experience. Staff should determine a range that seems to make sense, then look at the data.

Turning to the last example, let's say that after speaking with colleagues, staff determined that districts should be looked at further if they had any less than 50 percent for Indicator B5A. Staff should just go ahead and identify those cases.



D11 43.74%

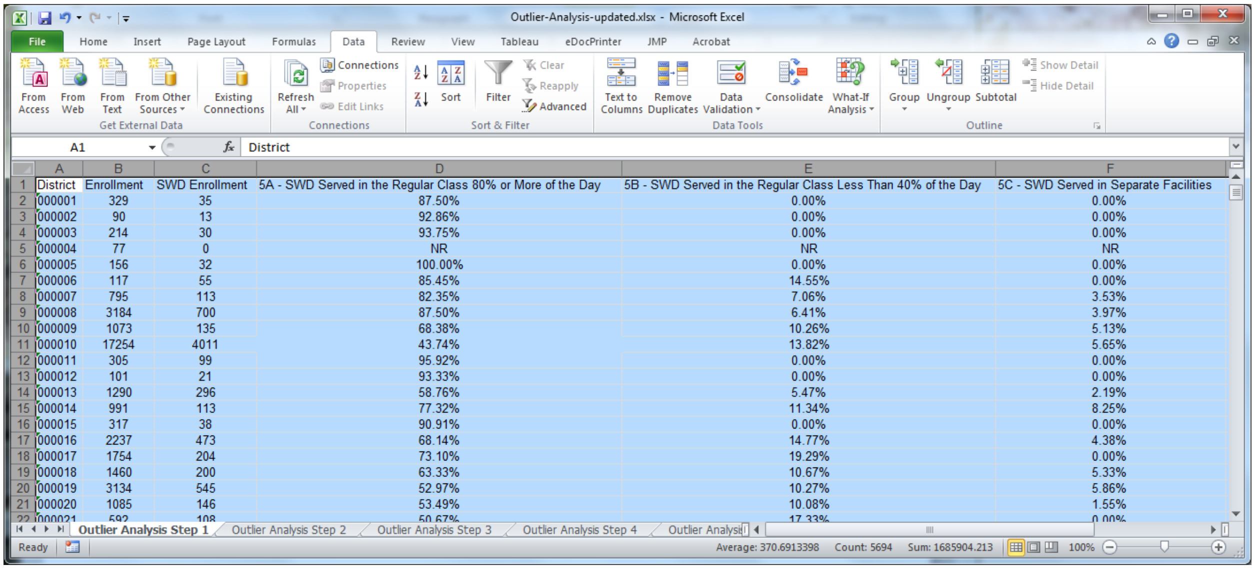
	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilities
2	000001	329	35	87.50%	0.00%	0.00%
3	000002	90	13	92.86%	0.00%	0.00%
4	000003	214	30	93.75%	0.00%	0.00%
5	000004	77	0	NR	NR	NR
6	000005	156	32	100.00%	0.00%	0.00%
7	000006	117	55	85.45%	14.55%	0.00%
8	000007	795	113	82.35%	7.06%	3.53%
9	000008	3184	700	87.50%	6.41%	3.97%
10	000009	1073	135	68.38%	10.26%	5.13%
11	000010	17254	4011	43.74%	13.82%	5.65%
12	000011	305	99	95.92%	0.00%	0.00%
13	000012	101	21	93.33%	0.00%	0.00%
14	000013	1290	296	58.76%	5.47%	2.19%
15	000014	991	113	77.32%	11.34%	8.25%
16	000015	317	38	90.91%	0.00%	0.00%
17	000016	2237	473	68.14%	14.77%	4.38%
18	000017	1754	204	73.10%	19.29%	0.00%
19	000018	1460	200	63.33%	10.67%	5.33%
20	000019	3134	545	52.97%	10.27%	5.86%
21	000020	1085	146	53.49%	10.08%	1.55%
22	000021	592	108	50.67%	17.33%	0.00%

Tutorial 3: Simply Sorting

Sometimes the easiest way to do an analysis would be to simply sort the data. The goal of an outlier analysis is not to support statistical analyses but, rather, to identify potential data issues. There is no harm in singling out data that staff would usually consider part of a normal range.

Step 1.

Staff should highlight their range of data.



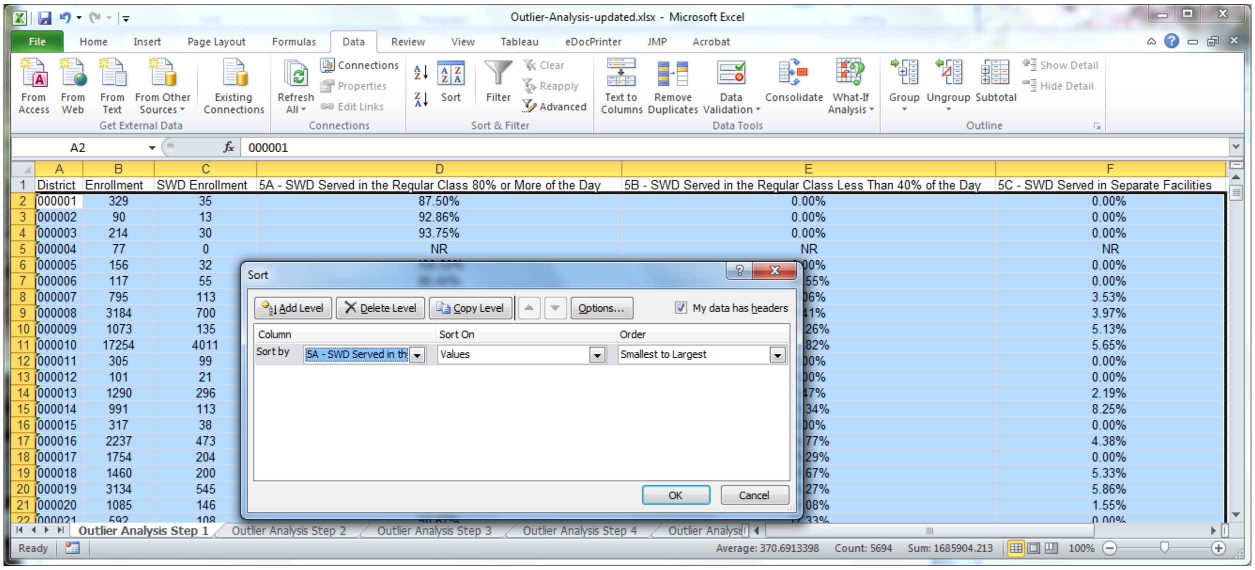
A1 | *fx* District

	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilities
2	000001	329	35	87.50%	0.00%	0.00%
3	000002	90	13	92.86%	0.00%	0.00%
4	000003	214	30	93.75%	0.00%	0.00%
5	000004	77	0	NR	NR	NR
6	000005	156	32	100.00%	0.00%	0.00%
7	000006	117	55	85.45%	14.55%	0.00%
8	000007	795	113	82.35%	7.06%	3.53%
9	000008	3184	700	87.50%	6.41%	3.97%
10	000009	1073	135	68.38%	10.26%	5.13%
11	000010	17254	4011	43.74%	13.82%	5.65%
12	000011	305	99	95.92%	0.00%	0.00%
13	000012	101	21	93.33%	0.00%	0.00%
14	000013	1290	296	58.76%	5.47%	2.19%
15	000014	991	113	77.32%	11.34%	8.25%
16	000015	317	38	90.91%	0.00%	0.00%
17	000016	2237	473	68.14%	14.77%	4.38%
18	000017	1754	204	73.10%	19.29%	0.00%
19	000018	1460	200	63.33%	10.67%	5.33%
20	000019	3134	545	52.97%	10.27%	5.86%
21	000020	1085	146	53.49%	10.08%	1.55%
22	000021	592	108	50.67%	17.33%	0.00%

Average: 370.6913398 Count: 5694 Sum: 1685904.213

Step 2.

Staff should sort by the measure.



A2 | *fx* 000001

	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilities
2	000001	329	35	87.50%	0.00%	0.00%
3	000002	90	13	92.86%	0.00%	0.00%
4	000003	214	30	93.75%	0.00%	0.00%
5	000004	77	0	NR	NR	NR
6	000005	156	32		00%	0.00%
7	000006	117	55		55%	0.00%
8	000007	795	113		06%	3.53%
9	000008	3184	700		41%	3.97%
10	000009	1073	135		26%	5.13%
11	000010	17254	4011		82%	5.65%
12	000011	305	99		00%	0.00%
13	000012	101	21		00%	0.00%
14	000013	1290	296		47%	2.19%
15	000014	991	113		34%	8.25%
16	000015	317	38		00%	0.00%
17	000016	2237	473		77%	4.38%
18	000017	1754	204		29%	0.00%
19	000018	1460	200		67%	5.33%
20	000019	3134	545		27%	5.86%
21	000020	1085	146		08%	1.55%
22	000021	592	108		33%	0.00%

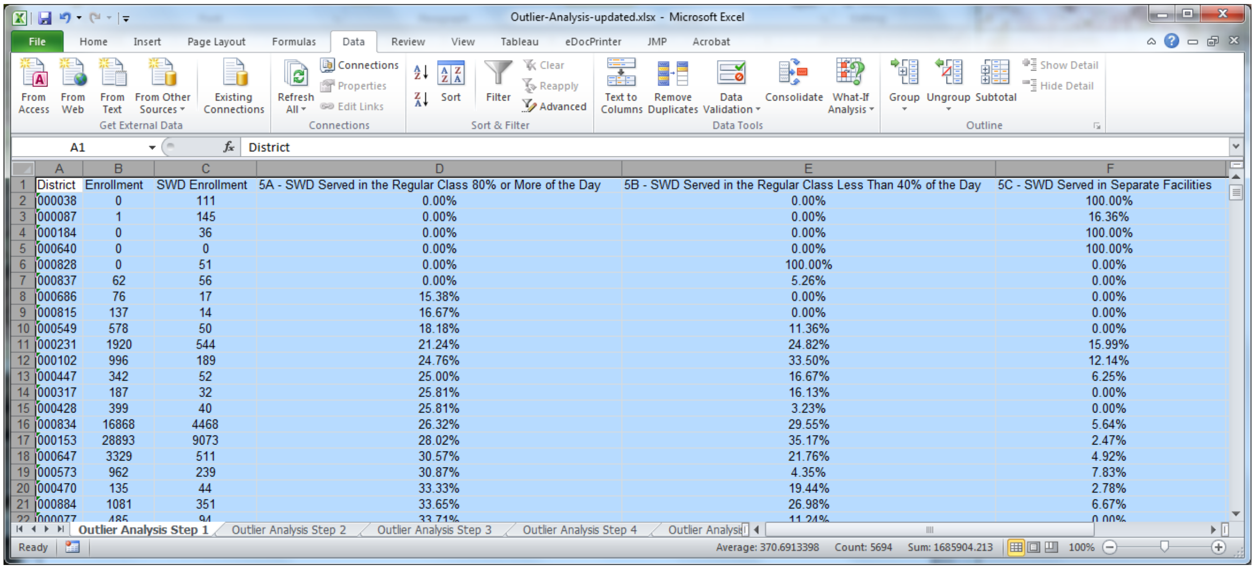
☑ My data has headers

Column	Sort On	Order
Sort by 5A - SWD Served in th	Values	Smallest to Largest

Average: 370.6913398 Count: 5694 Sum: 1685904.213

Step 3.

Staff should look deeper into the values at the top and bottom of the sort.



	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilities
2	000038	0	111	0.00%	0.00%	100.00%
3	000087	1	145	0.00%	0.00%	16.36%
4	000184	0	36	0.00%	0.00%	100.00%
5	000640	0	0	0.00%	0.00%	100.00%
6	000828	0	51	0.00%	100.00%	0.00%
7	000837	62	56	0.00%	5.26%	0.00%
8	000686	76	17	15.38%	0.00%	0.00%
9	000815	137	14	16.67%	0.00%	0.00%
10	000549	578	50	18.18%	11.36%	0.00%
11	000231	1920	544	21.24%	24.82%	15.99%
12	000102	996	189	24.76%	33.50%	12.14%
13	000447	342	52	25.00%	16.67%	6.25%
14	000317	187	32	25.81%	16.13%	0.00%
15	000428	399	40	25.81%	3.23%	0.00%
16	000834	16868	4468	26.32%	29.55%	5.64%
17	000153	28893	9073	28.02%	35.17%	2.47%
18	000647	3329	511	30.57%	21.76%	4.92%
19	000573	962	239	30.87%	4.35%	7.83%
20	000470	135	44	33.33%	19.44%	2.78%
21	000884	1081	351	33.65%	26.98%	6.67%
22	000077	486	94	33.71%	11.24%	0.00%

Data Visualization Support for Outlier Analyses

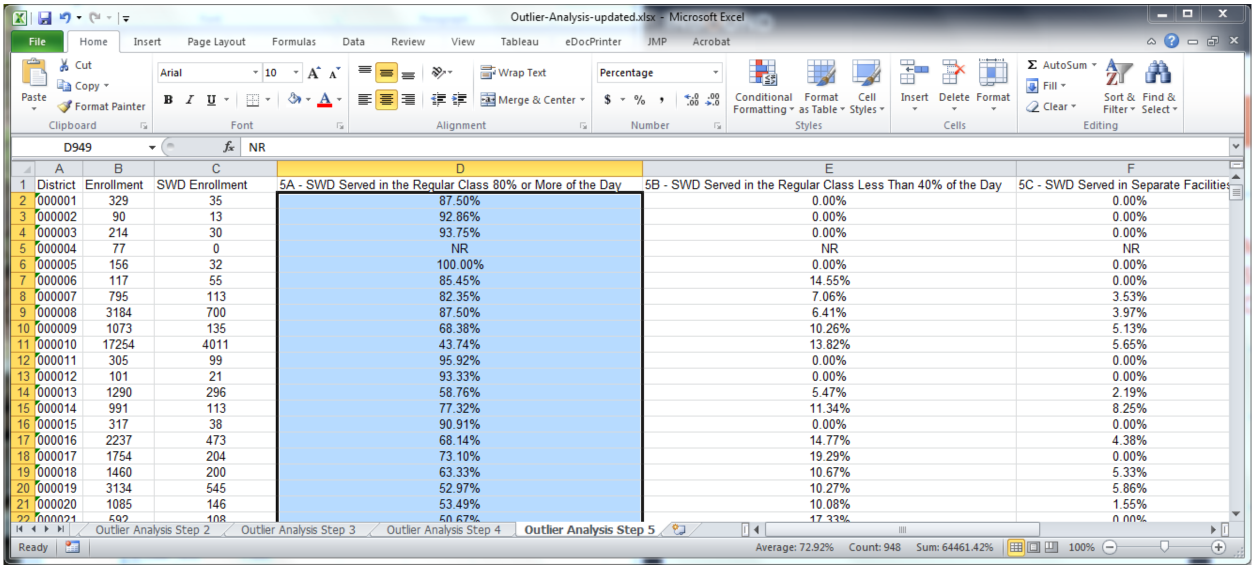
Data visualization can help support outlier analyses by making identified outliers more noticeable. Here are a few ways state staff can use data visualization to support their outlier analysis.

Tutorial 4: Heat Maps in Excel

By using conditional formatting in conjunction with the approaches shown previously, staff can make outliers stand out. This can be particularly important when there are multiple columns of data or many rows.

Step 1.

Staff should highlight the range of data and click on the *Conditional Formatting* button in the *Home* tab.



D949 NR

	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Served in the Regular Class Less Than 40% of the Day	5C - SWD Served in Separate Facilities
2	000001	329	35	87.50%	0.00%	0.00%
3	000002	90	13	92.86%	0.00%	0.00%
4	000003	214	30	93.75%	0.00%	0.00%
5	000004	77	0	NR	NR	NR
6	000005	156	32	100.00%	0.00%	0.00%
7	000006	117	55	85.45%	14.55%	0.00%
8	000007	795	113	82.35%	7.06%	3.53%
9	000008	3184	700	87.50%	6.41%	3.97%
10	000009	1073	135	68.38%	10.26%	5.13%
11	000010	17254	4011	43.74%	13.82%	5.65%
12	000011	305	99	95.92%	0.00%	0.00%
13	000012	101	21	93.33%	0.00%	0.00%
14	000013	1290	296	58.76%	5.47%	2.19%
15	000014	991	113	77.32%	11.34%	8.25%
16	000015	317	38	90.91%	0.00%	0.00%
17	000016	2237	473	68.14%	14.77%	4.38%
18	000017	1754	204	73.10%	19.29%	0.00%
19	000018	1460	200	63.33%	10.67%	5.33%
20	000019	3134	545	52.97%	10.27%	5.86%
21	000020	1085	146	53.49%	10.08%	1.55%
22	000021	592	108	50.67%	17.33%	0.00%

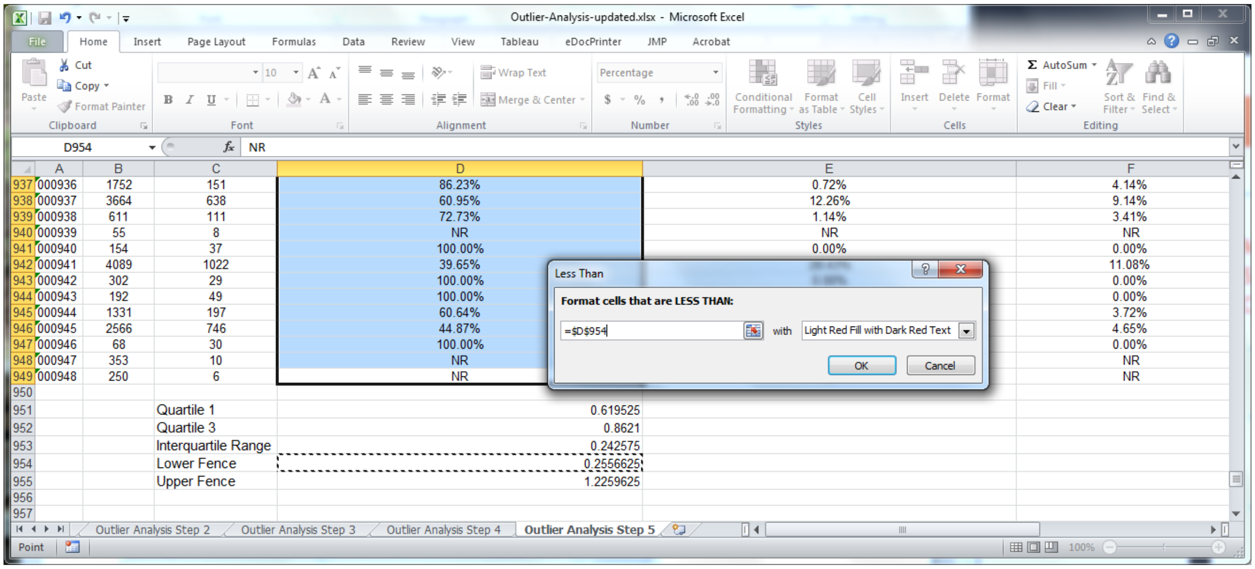
Outlier Analysis Step 2 | Outlier Analysis Step 3 | Outlier Analysis Step 4 | **Outlier Analysis Step 5**

Ready Average: 72.92% Count: 948 Sum: 64461.42% 100%

Step 2.

If staff determined the normal range w by using the interquartile range, they should select *Highlight Cells Rules*, then select *Less Than*. In the box that asks for a number, they should enter the cell location of the Lower Fence value. They can also adjust the format (color/font/additional symbols).

Staff should do the same for the Upper Fence but use *Greater Than*.



D954 | NR

	A	B	C	D	E	F
937	000936	1752	151	86.23%	0.72%	4.14%
938	000937	3664	638	60.95%	12.26%	9.14%
939	000938	611	111	72.73%	1.14%	3.41%
940	000939	55	8	NR	NR	NR
941	000940	154	37	100.00%	0.00%	0.00%
942	000941	4089	1022	39.65%		11.08%
943	000942	302	29	100.00%		0.00%
944	000943	192	49	100.00%		0.00%
945	000944	1331	197	60.64%		3.72%
946	000945	2566	746	44.87%		4.65%
947	000946	68	30	100.00%		0.00%
948	000947	353	10	NR		NR
949	000948	250	6	NR		NR
950						
951			Quartile 1	0.619525		
952			Quartile 3	0.8621		
953			Interquartile Range	0.242575		
954			Lower Fence	0.2556625		
955			Upper Fence	1.2259625		
956						
957						

Less Than

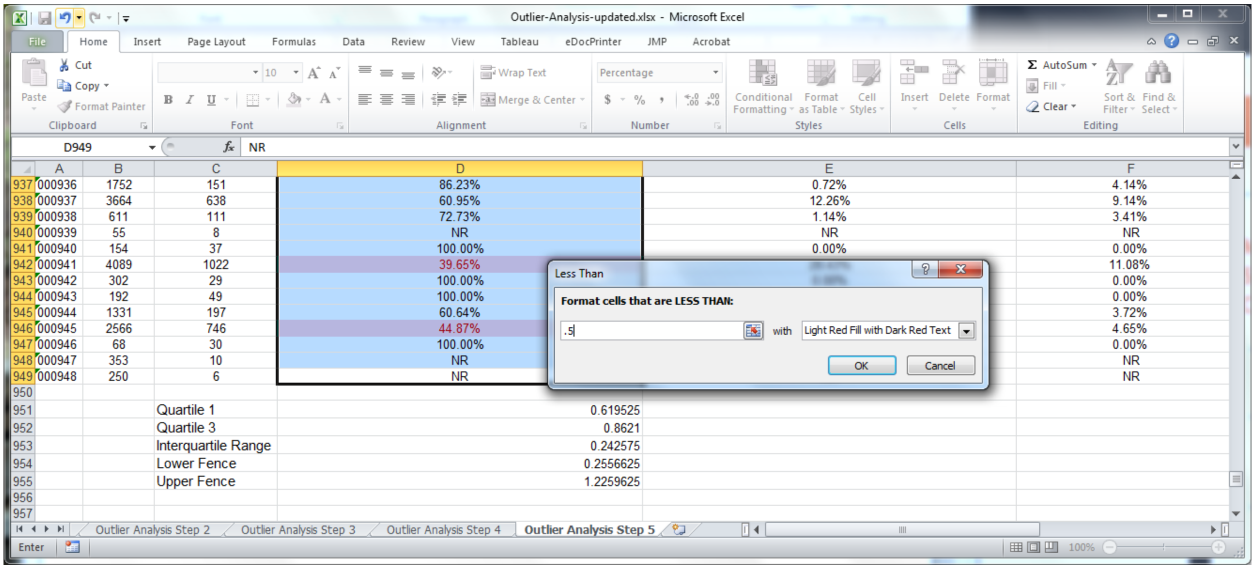
Format cells that are LESS THAN:

=D954 with Light Red Fill with Dark Red Text

OK Cancel

Step 3.

If staff determined the normal range using a qualitative approach, they should follow the same instructions as in Step 2. They should either enter the number directly into the *LESS THAN* box or place the Lower Fence value into a cell in the worksheet and refer to that.



D949 | NR

	A	B	C	D	E	F
937	000936	1752	151	86.23%	0.72%	4.14%
938	000937	3664	638	60.95%	12.26%	9.14%
939	000938	611	111	72.73%	1.14%	3.41%
940	000939	55	8	NR	NR	NR
941	000940	154	37	100.00%	0.00%	0.00%
942	000941	4089	1022	39.65%		11.08%
943	000942	302	29	100.00%		0.00%
944	000943	192	49	100.00%		0.00%
945	000944	1331	197	60.64%		3.72%
946	000945	2566	746	44.87%		4.65%
947	000946	68	30	100.00%		0.00%
948	000947	353	10	NR		NR
949	000948	250	6	NR		NR
950						
951			Quartile 1	0.619525		
952			Quartile 3	0.8621		
953			Interquartile Range	0.242575		
954			Lower Fence	0.2556625		
955			Upper Fence	1.2259625		
956						
957						

Less Than

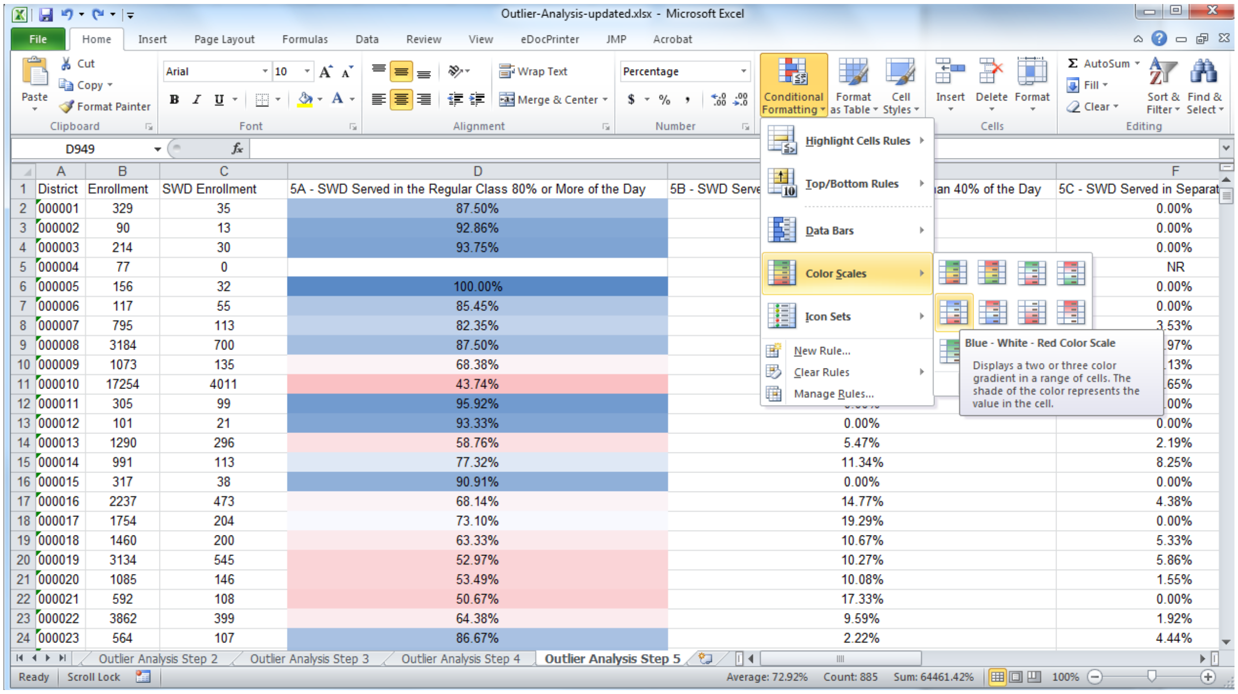
Format cells that are LESS THAN:

.5 with Light Red Fill with Dark Red Text

OK Cancel

Alternative Quick Heat Maps

A quicker approach to heat maps in Excel is using the *Color Scales* feature found under the *Conditional Formatting* drop-down menu. This will automatically color a set of selected cells based on the range of values. Staff also can use data bars and icon sets to quickly identify possible outliers.



Blue - White - Red Color Scale

Displays a two or three color gradient in a range of cells. The shade of the color represents the value in the cell.

	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day	5B - SWD Serve an 40% of the Day	5C - SWD Served in Separat
2	000001	329	35	87.50%		0.00%
3	000002	90	13	92.86%		0.00%
4	000003	214	30	93.75%		0.00%
5	000004	77	0			NR
6	000005	156	32	100.00%		0.00%
7	000006	117	55	85.45%		0.00%
8	000007	795	113	82.35%		3.53%
9	000008	3184	700	87.50%		.97%
10	000009	1073	135	68.38%		.13%
11	000010	17254	4011	43.74%		.65%
12	000011	305	99	95.92%		.00%
13	000012	101	21	93.33%	0.00%	0.00%
14	000013	1290	296	58.76%	5.47%	2.19%
15	000014	991	113	77.32%	11.34%	8.25%
16	000015	317	38	90.91%	0.00%	0.00%
17	000016	2237	473	68.14%	14.77%	4.38%
18	000017	1754	204	73.10%	19.29%	0.00%
19	000018	1460	200	63.33%	10.67%	5.33%
20	000019	3134	545	52.97%	10.27%	5.86%
21	000020	1085	146	53.49%	10.08%	1.55%
22	000021	592	108	50.67%	17.33%	0.00%
23	000022	3862	399	64.38%	9.59%	1.92%
24	000023	564	107	86.67%	2.22%	4.44%

Average: 72.92% Count: 885 Sum: 64461.42%

Tutorial 5: Dot Plots in Excel

Another approach to identifying outliers involves visualizing the data. An easy approach is using an in-cell formula that will create a simple dot plot next to the data. Once the formula creates the plot, staff simply look for data points that do not seem to fit with the others.

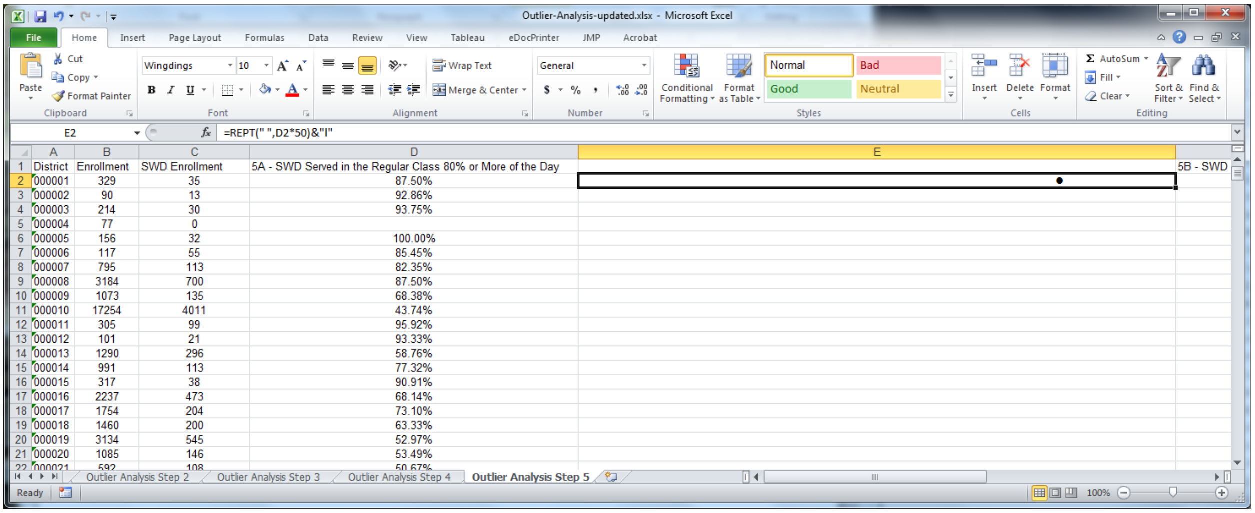
By using an Excel formula to create an in-cell chart, dot plots will always remain in line with the data. Excel's standard chart functions also can assist in identifying outliers but the standard chart functions will ultimately be disconnected from the original data.

Step 1.

This example uses Excel's REPT function to create this visual. Staff should start by selecting a cell next to the first data point. Basically, staff is instructing Excel to put in a series of repeating blanks based on the data and then adding some kind of character at the end.

In this example, the formula looks like this:
=REPT (" ",B2*50)&"l"

In order to get the dot at the end, staff should use the Wingdings font. The *50 multiplier creates the spacing. If staff desire less spacing, they should use a smaller multiplier.

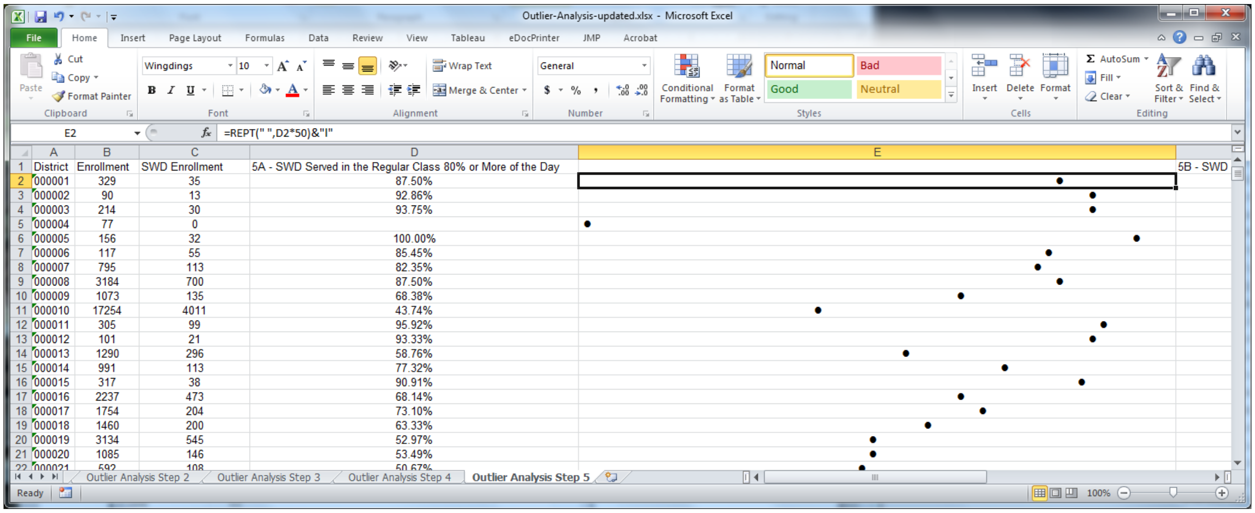


E2 =REPT(" ",D2*50)&"l"

	A	B	C	D	E	F
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day		5B - SWD
2	000001	329	35	87.50%	●	
3	000002	90	13	92.86%		
4	000003	214	30	93.75%		
5	000004	77	0			
6	000005	156	32	100.00%		
7	000006	117	55	85.45%		
8	000007	795	113	82.35%		
9	000008	3184	700	87.50%		
10	000009	1073	135	68.38%		
11	000010	17254	4011	43.74%		
12	000011	305	99	95.92%		
13	000012	101	21	93.33%		
14	000013	1290	296	58.76%		
15	000014	991	113	77.32%		
16	000015	317	38	90.91%		
17	000016	2237	473	68.14%		
18	000017	1754	204	73.10%		
19	000018	1460	200	63.33%		
20	000019	3134	545	52.97%		
21	000020	1085	146	53.49%		
22	000021	592	108	50.67%		

Step 2.

Staff then just need to copy the formula down once they are happy with the look of the first cell. That is all that it takes!



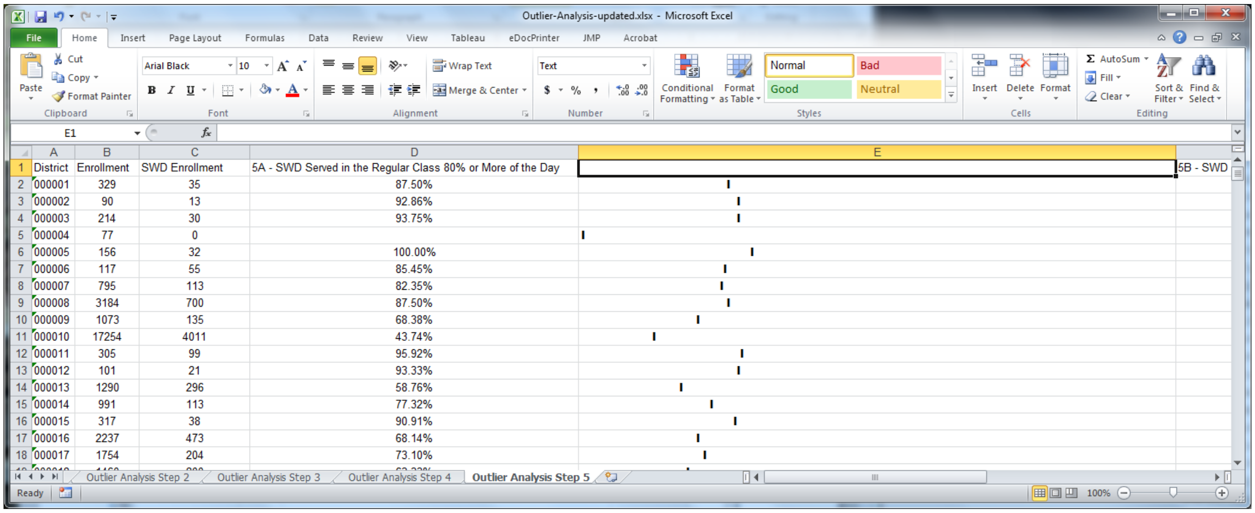
Font: Wingdings, 10; Number format: General

E2 =REPT(" ",D2*50)&"l"

	A	B	C	D	E	
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day		5B - SWD
2	000001	329	35	87.50%	●	
3	000002	90	13	92.86%	●	
4	000003	214	30	93.75%	●	
5	000004	77	0		●	
6	000005	156	32	100.00%	●	
7	000006	117	55	85.45%	●	
8	000007	795	113	82.35%	●	
9	000008	3184	700	87.50%	●	
10	000009	1073	135	68.38%	●	
11	000010	17254	4011	43.74%	●	
12	000011	305	99	95.92%	●	
13	000012	101	21	93.33%	●	
14	000013	1290	296	58.76%	●	
15	000014	991	113	77.32%	●	
16	000015	317	38	90.91%	●	
17	000016	2237	473	68.14%	●	
18	000017	1754	204	73.10%	●	
19	000018	1460	200	63.33%	●	
20	000019	3134	545	52.97%	●	
21	000020	1085	146	53.49%	●	
22	000021	592	108	50.67%	●	

Step 3.

It is easy to change the font or change the character and get a different look. Here is an alternative using the same exact formula with the Arial Black font in 10 point instead of Wingdings. The font change also changes the spacing, so staff should set up cells within the same column using the same font.



Font: Arial Black, 10; Number format: Text

E1

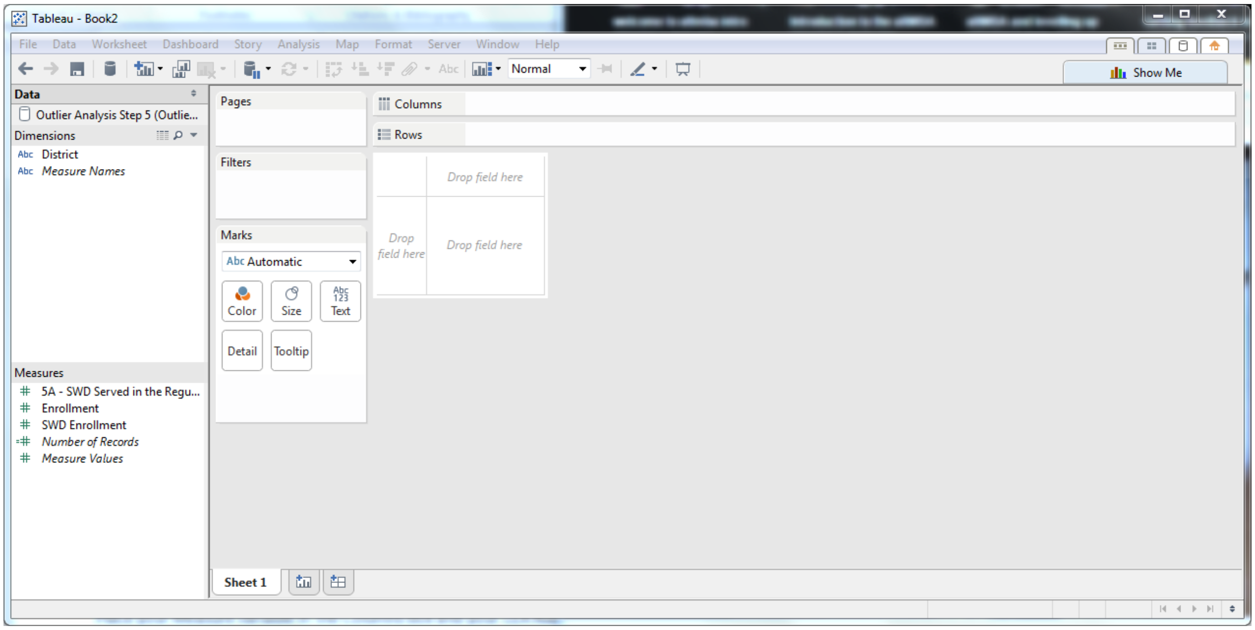
	A	B	C	D	E	
1	District	Enrollment	SWD Enrollment	5A - SWD Served in the Regular Class 80% or More of the Day		5B - SWD
2	000001	329	35	87.50%	I	
3	000002	90	13	92.86%	I	
4	000003	214	30	93.75%	I	
5	000004	77	0		I	
6	000005	156	32	100.00%	I	
7	000006	117	55	85.45%	I	
8	000007	795	113	82.35%	I	
9	000008	3184	700	87.50%	I	
10	000009	1073	135	68.38%	I	
11	000010	17254	4011	43.74%	I	
12	000011	305	99	95.92%	I	
13	000012	101	21	93.33%	I	
14	000013	1290	296	58.76%	I	
15	000014	991	113	77.32%	I	
16	000015	317	38	90.91%	I	
17	000016	2237	473	68.14%	I	
18	000017	1754	204	73.10%	I	

Tutorial 6: Dot Plots in Tableau

Using an interactive visualization program like Tableau can be useful if there is a large amount of data. Such a program allows the user to quickly and easily visualize hundreds of rows and multiple measures. If visualizing public data, state staff can use Tableau Public for free. For private data, staff would need at least a personal version of the Tableau desktop license.

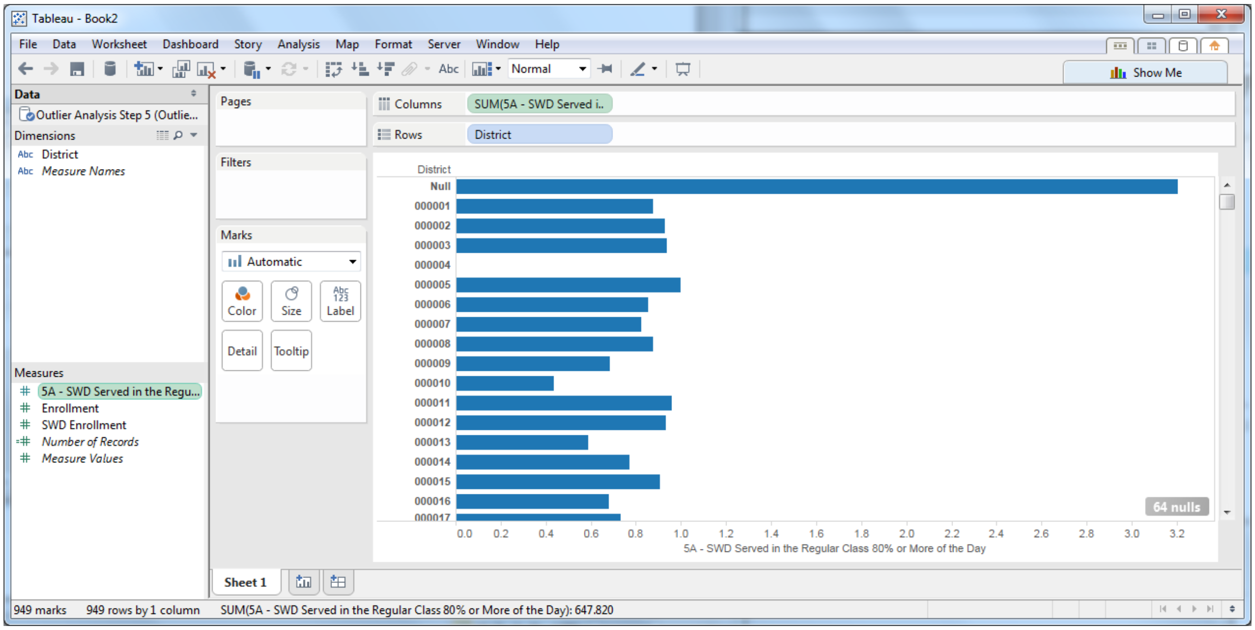
Step 1.

Staff should open the Excel data in Tableau. The measures staff use should be under the *Measures* space in Tableau, and descriptive variables, such as district, should be under *Dimensions*. If this is not the case, staff should right-click on the variable and convert.



Step 2.

Staff should place the measure variable in the *Columns* box and the district variable in the *Rows* box. The default should be a bar graph. If satisfied with the bar chart, staff stop here.



Columns: SUM(5A - SWD Served i..

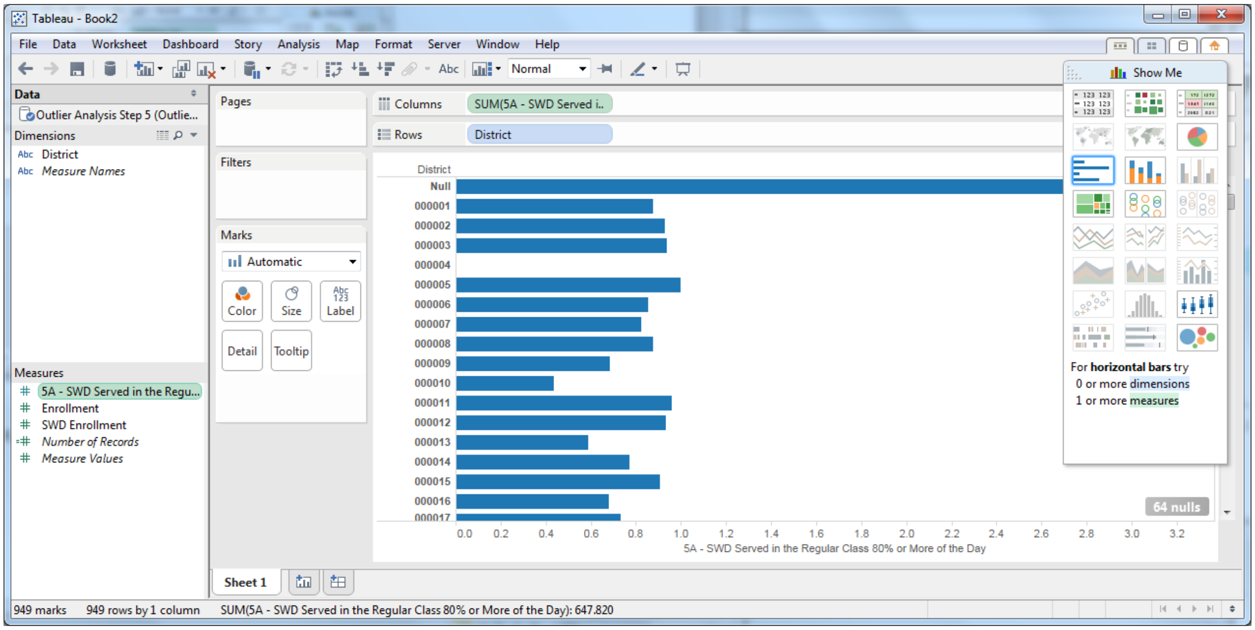
Rows: District

Normal

949 marks 949 rows by 1 column SUM(5A - SWD Served in the Regular Class 80% or More of the Day): 647.820

Step 3.

Click on the *Show Me* tab and select the bar chart.



Columns: SUM(5A - SWD Served i..

Rows: District

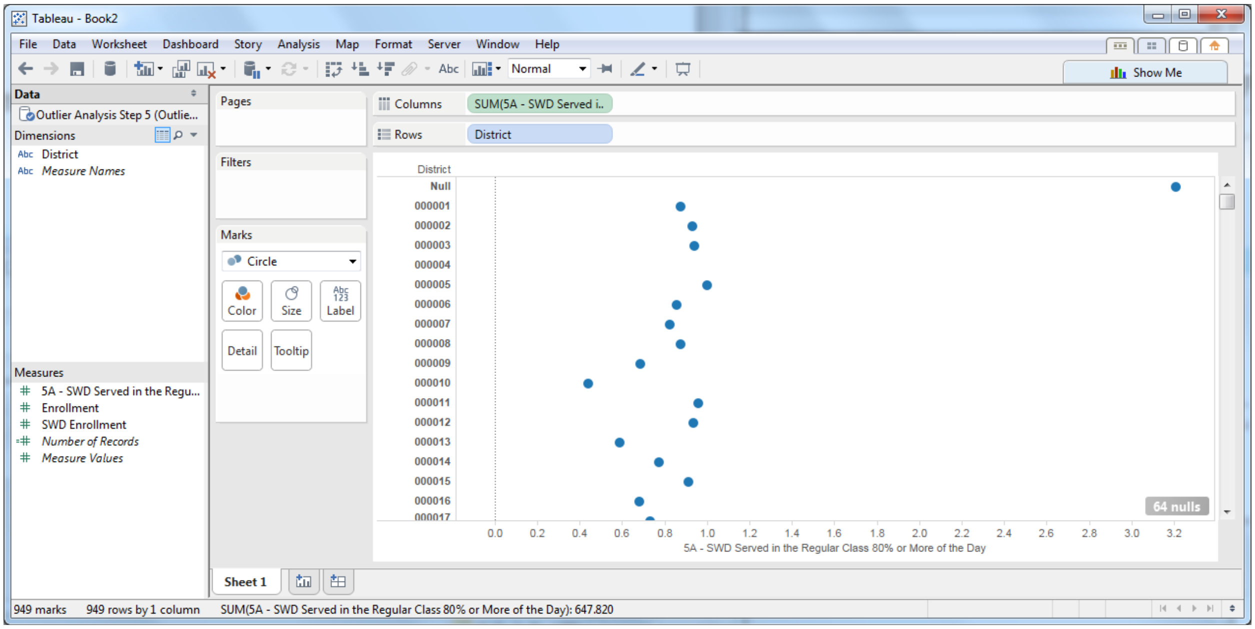
Normal

For **horizontal bars** try
0 or more dimensions
1 or more measures

949 marks 949 rows by 1 column SUM(5A - SWD Served in the Regular Class 80% or More of the Day): 647.820

Step 4.

To change to a dot plot, staff should just change the *Marks* from Automatic to Circle.



Marks: Circle

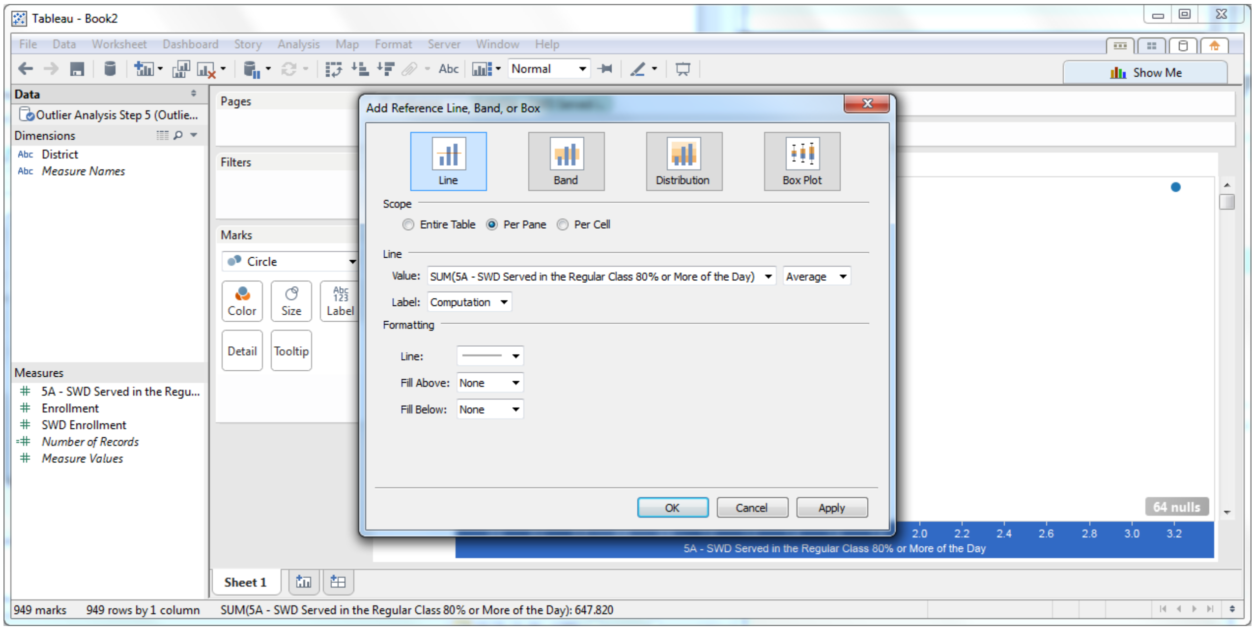
Columns: SUM(5A - SWD Served i..

Rows: District

949 marks 949 rows by 1 column SUM(5A - SWD Served in the Regular Class 80% or More of the Day): 647.820

Step 5.

To create a line to identify the Lower and Upper Fences, staff should just right-click on the x axis and select *Add Reference Line, Band, or Box*. Then in the *Value* box, select *Create New Parameter*.



Scope: ○ Entire Table ● Per Pane ○ Per Cell

Line

Value: SUM(5A - SWD Served in the Regular Class 80% or More of the Day) Average

Label: Computation

Formatting

Line:

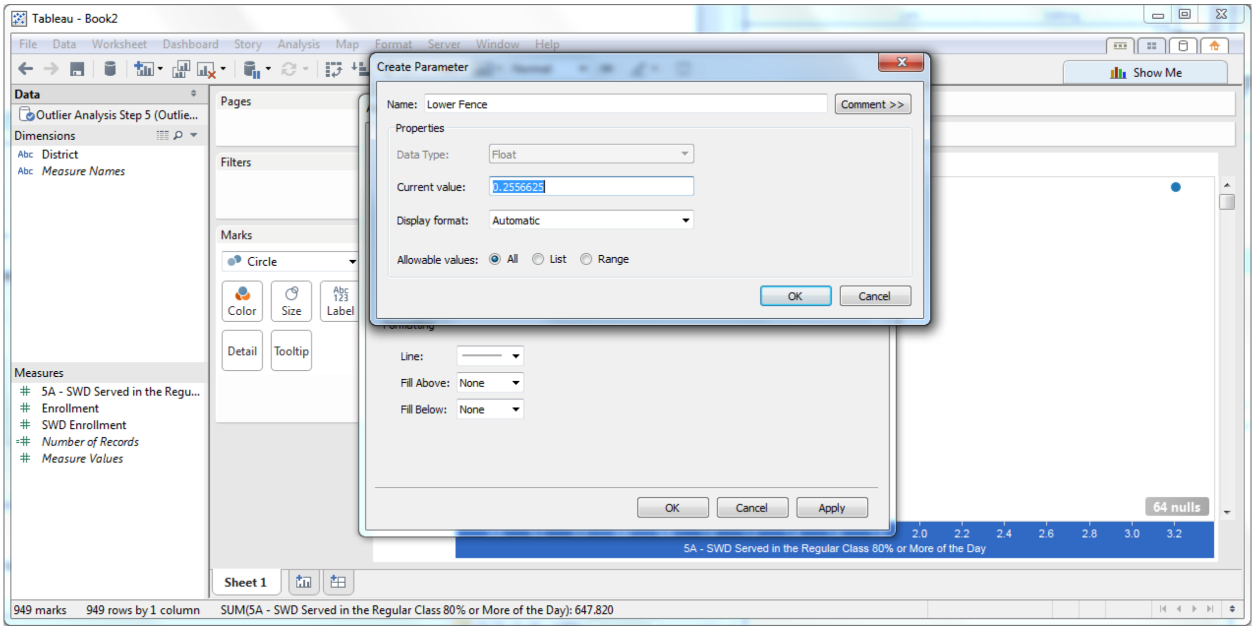
Fill Above: None

Fill Below: None

949 marks 949 rows by 1 column SUM(5A - SWD Served in the Regular Class 80% or More of the Day): 647.820

Step 6.

Staff should change the name of the parameter to “Fence” and set the current value as the Lower Fence Value and click *OK*. Then, staff should click *Apply* and *OK*.



Create Parameter

Name: Lower Fence

Properties

Data Type: Float

Current value: 0.2556625

Display format: Automatic

Allowable values: ● All ○ List ○ Range

Line:

Fill Above: None

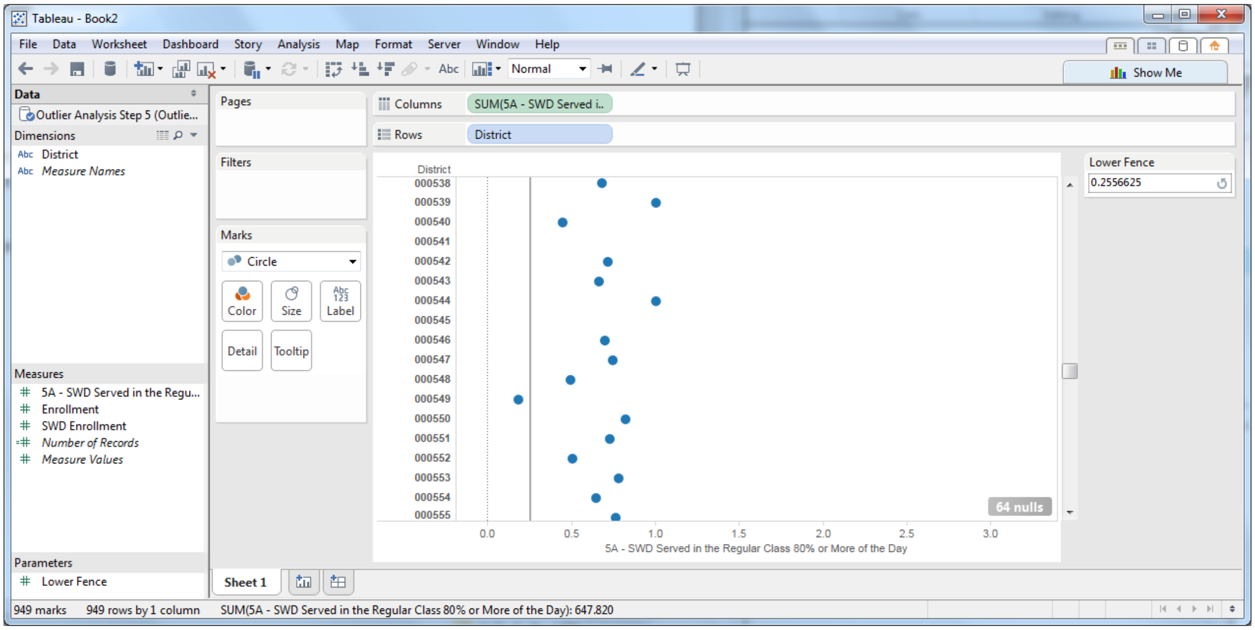
Fill Below: None

949 marks 949 rows by 1 column SUM(5A - SWD Served in the Regular Class 80% or More of the Day): 647.820

Step 7.

To include both the Upper and Lower Fences, staff should just follow the same steps a second time.

Staff may also opt to use the *Parameters* box that the program created on the screen to shift the Fence based on what they are trying to see.



Columns: SUM(5A - SWD Served i..

Rows: District

Lower Fence: 0.2556625

949 marks 949 rows by 1 column SUM(5A - SWD Served in the Regular Class 80% or More of the Day): 647.820

Conclusion

This guide includes a handful of approaches state agency staff can use to identify and visualize outliers. There are many others. Staff should find a set of approaches that works well for their state and their data and then apply the approaches systematically.

State staff should visit ideadata.org to get in touch with their IDC State Liaison if they have any questions about these approaches.