

|  |
| --- |
|  |
| City of Dallas - Water Quality Analysis |
| Jan – Apr 2020  SMU MS Business Analytics  Authored by: Emily Fogg, Hannah Roark, and Hazel Wu |



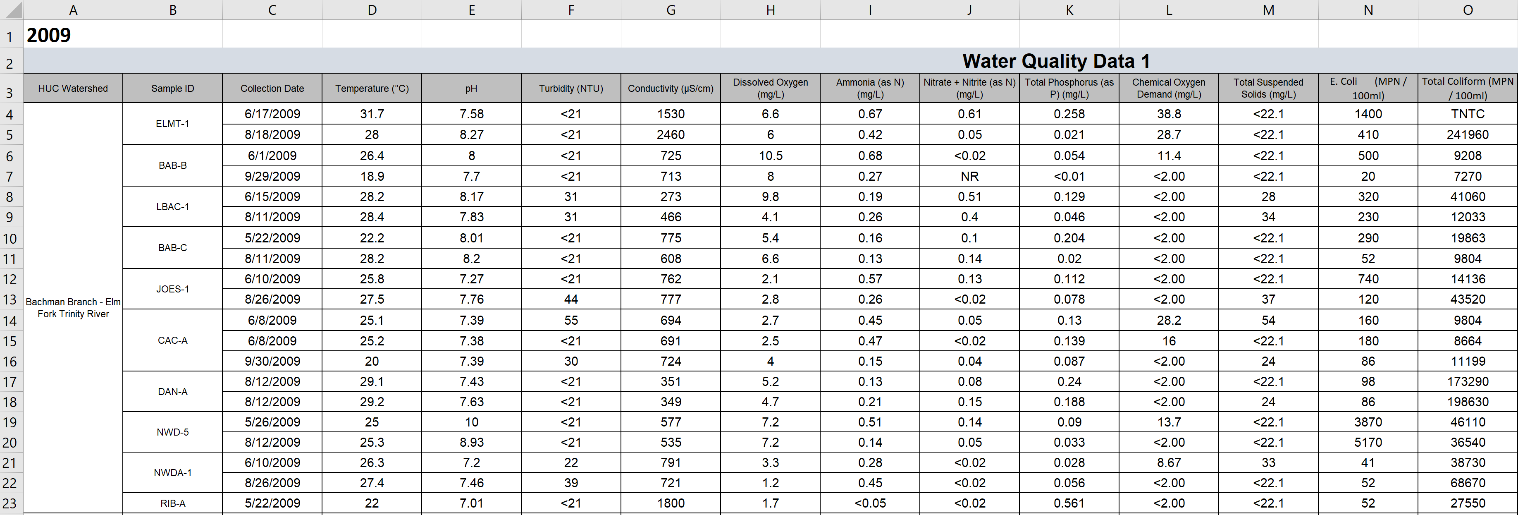
# Excel Data Preparation Steps

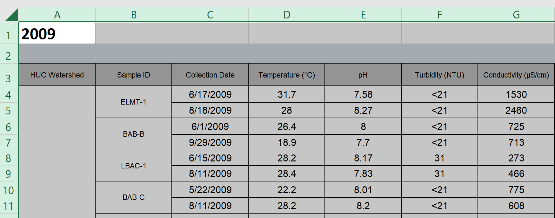
Version: Microsoft Excel for Office 365 MSO (16.0.11929.20618) 64-bit

Introduction:

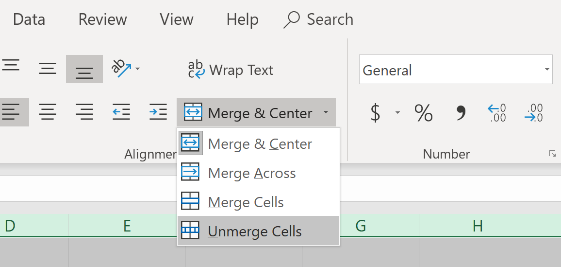
The raw excel file contains Water Quality data spanning 18 columns. The data needs to be configured to a specific format to enable statistical analysis and visualization in R and Tableau, respectively. The following steps will resolve the issues with:

* Merged cells
* Empty Rows
* Inconsistent data types
* Inconsistent syntax
* Common human errors created at the data entry step

Figure Raw Data

Step 1 – Unmerge Cells

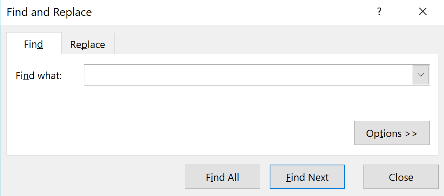
Select all cells by clicking the triangle in the top left corner.



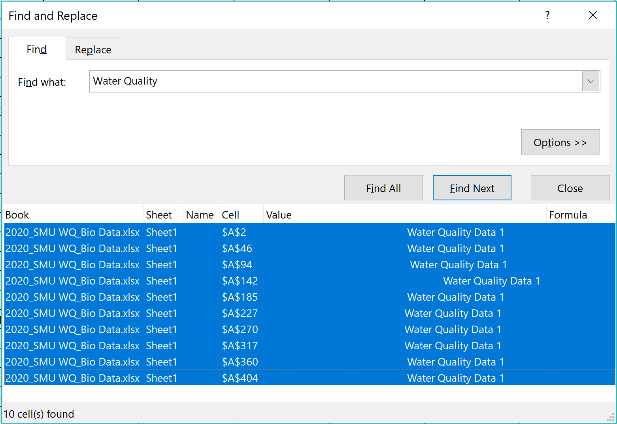
Click “Merge & Center” and then click “Unmerge Cells”

\*\* Do not worry about the empty cells at this stage.

Step 2 – Delete Redundant Headers

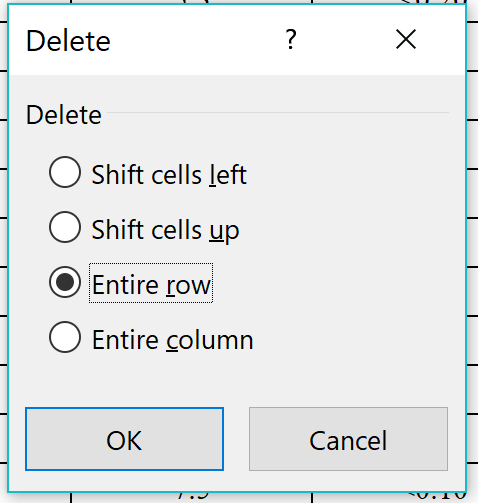
Select Column A and press CTRL + F. This will open the Find and Replace dialogue box pictured below.

In the box, type “Water Quality” and click “Find All.” 10 values should appear (1 for each year).



Select the first entry. The row should highlight blue. Hold down the SHIFT key and select the last entry. All of the entries should now be highlighted blue. Click the “X” in the top right corner to close the dialogue box.

Before clicking on anything else, press CTRL + - (control plus the minus sign) to open the Delete dialogue box.



Select *Entire Row* and click “OK.”

You have now successfully deleted the redundant header “Water Quality Data 1”

Step 3 – Delete Redundant Year Headers

Repeat the process from step 2 except instead of typing “Water Quality” into the dialogue box, type “20.” This will pull up all 10 years at once since they all begin with 20. Delete these rows.

Step 4 – Delete Empty Rows

Fortunately, there are no missing values in Column C (the date column). We will use this to our advantage to locate the empty rows that are used to visually separate data between different years.

Repeat the process from step 2, but instead of typing “Water Quality” into the dialogue box, leave it empty. It is essential to perform this step on Column C. Performing this step on Column A will delete almost all the data.

Step 5 – Confirm Column Consistency

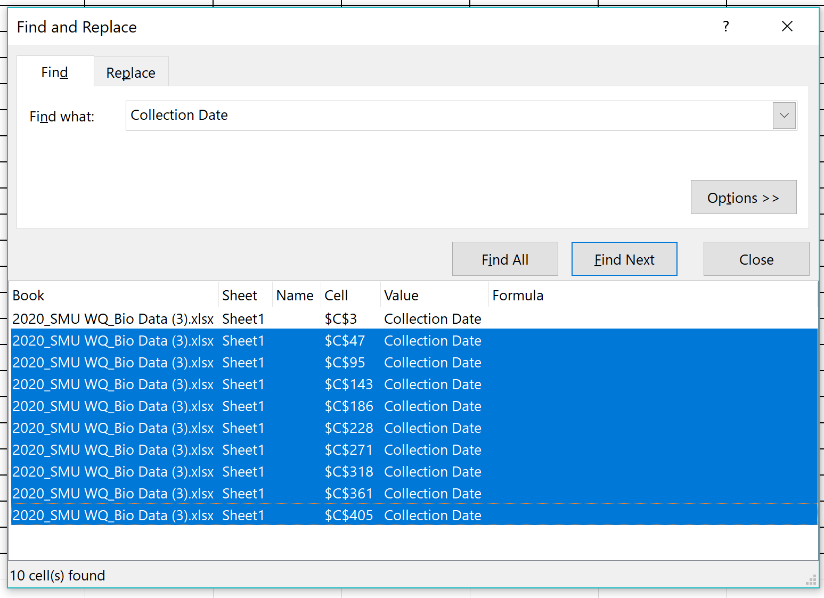
In the raw data file, column headers are repeated for each year. However, there is no guarantee that the columns are arranged in the same order for every year. First, we need to confirm the data is situated in the right column. Then, we need to delete the duplicate headers.

Column Consistency

First, copy the column header from column A (should be cell A1 – HUC Watershed). Then, select Column A and use CTRL + F to find all instances of “HUC Watershed.” There should be 10 (1 for each year). Repeat this step for all columns. Each time, confirming there are 10 values. If there are less than 10 values, then that means a column is swapped somewhere in the file.

Delete Redundant Headers

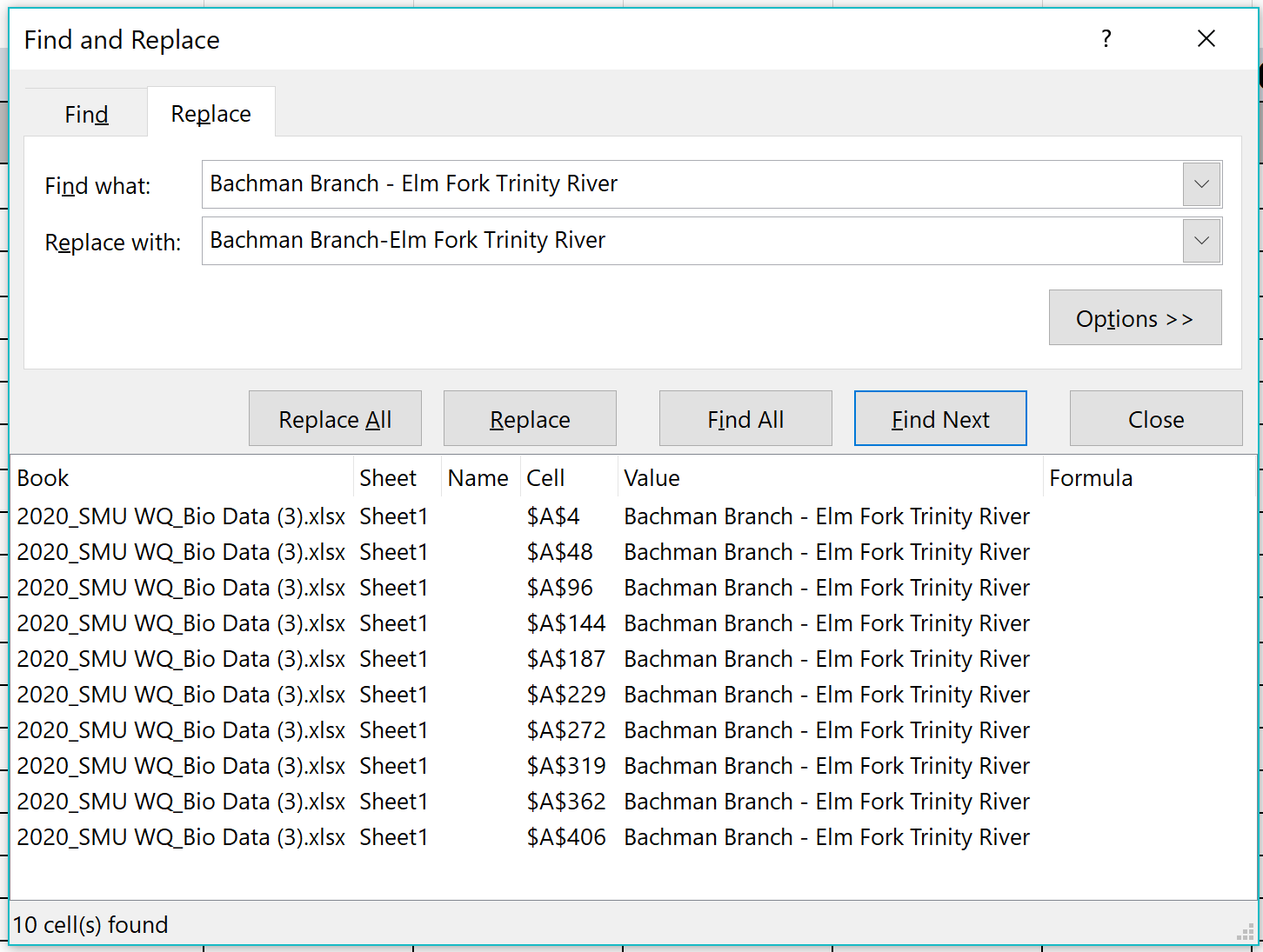
Repeat the process from step 4 to delete the redundant headers. Select Column C and use CTRL + F to find all instances of “Collection Date.” However, instead of using CTRL + SHIFT to select all rows, we need to select all rows *except the first row*. Close the dialogue box and use CTRL + - to delete the redundant headers.



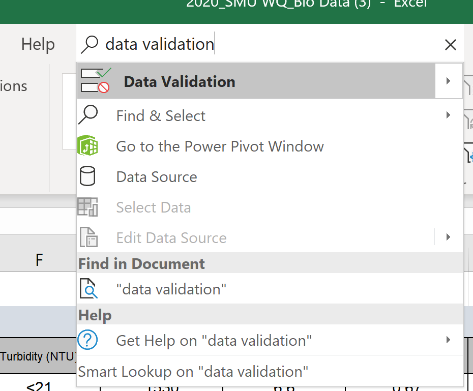
Keep the First Instance to preserve the header at the top

Step 6 – Confirm Watershed Name Consistency

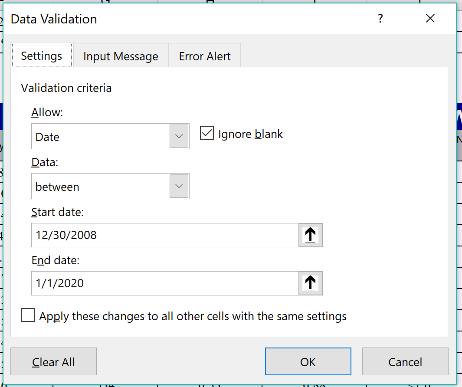
First, copy the contents of the cell containing the watershed name. Next, select Column A and use CTRL + F to open the Find and Replace dialogue box. Paste the watershed name into the text box. Select “Find All.”

Finally, click on the “Replace” tab. Insert the desired text in the *Replace With* text box. In the example below, the spaces between the dash in Bachman Branch – Elm Fork Trinity River are removed to ensure uniformity.

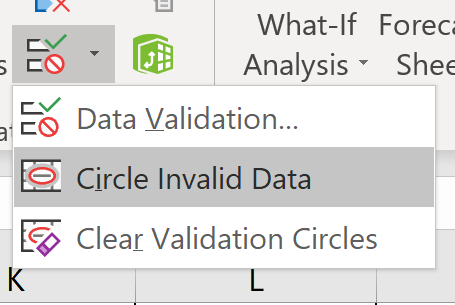
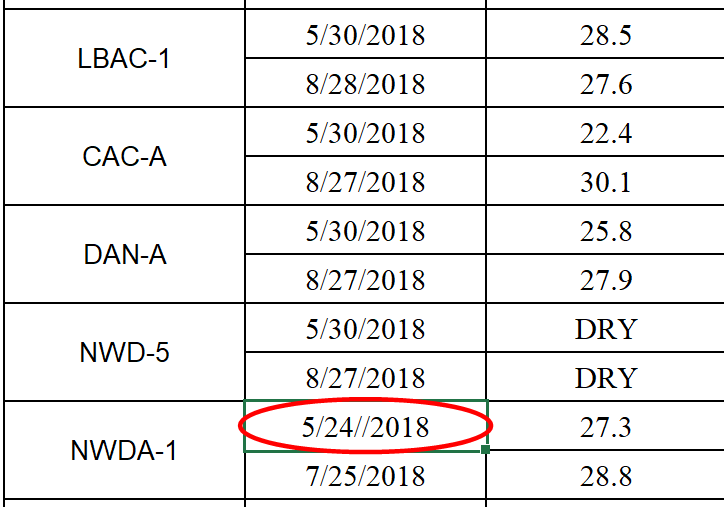
Step 7 – Data Validation to Enforce Uniformity



This step will help to resolve some of the human errors created at the data entry stage. First, use Excel’s search feature to locate “Data Validation.” Open the tool.



Configure the dialogue box using the values pictured left.

Finally, use the “Circle Invalid Data” feature to circle all instances that do not meet the specified format.

In the image picture left, the data validation circle was used to locate a date that accidentally had two slash marks.

Step 8 – Delete Potentially Extraneous Columns

As a final step, select the next 8 empty columns (S-Z) . Right click and select “Delete.” This will ensure that any accidental spaces, commas, or notes entered in random cells are not accidentally recognized by R as a data points.

After completing these steps, the file is ready to load into R.

# Tableau

Version: Tableau Desktop 2020.1.2 64 bit

Download the free public app [here](https://public.tableau.com/en-us/s/). Or download a free two-week trial with editor capabilities [here](https://www.tableau.com/trial/download-tableau?utm_campaign_id=2017049&utm_campaign=Prospecting-CORE-ALL-ALL-ALL-ALL&utm_medium=Paid+Search&utm_source=Google+Search&utm_language=EN&utm_country=USCA&kw=tableau%20download&adgroup=CTX-Brand-Download-E&adused=ETA&matchtype=e&placement=&gclid=EAIaIQobChMIzr-0uqHn6AIVhYbACh0jxAAXEAAYASAAEgKY0fD_BwE&gclsrc=aw.ds).

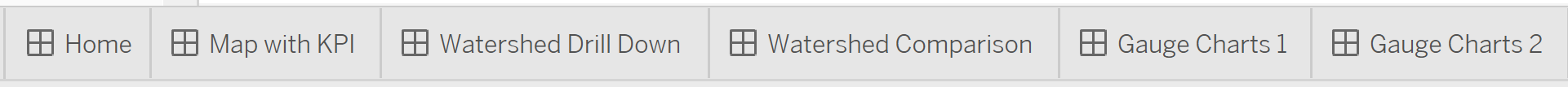
Introduction:

There are 6 Key Dashboards built to provide a Macro-, Meso-, and Micro-level analysis.

Home – Contains a navigable homepage. Clicking on an image routes the user to the desired dashboard.

Map with KPI – Includes a GIS map with polygon outlines of each watershed and a KPI chart that updates as the user hovers his or her mouse on a specific watershed. The KPI dots provide a cursory, yet dynamic view of the most recent health of a given watershed.

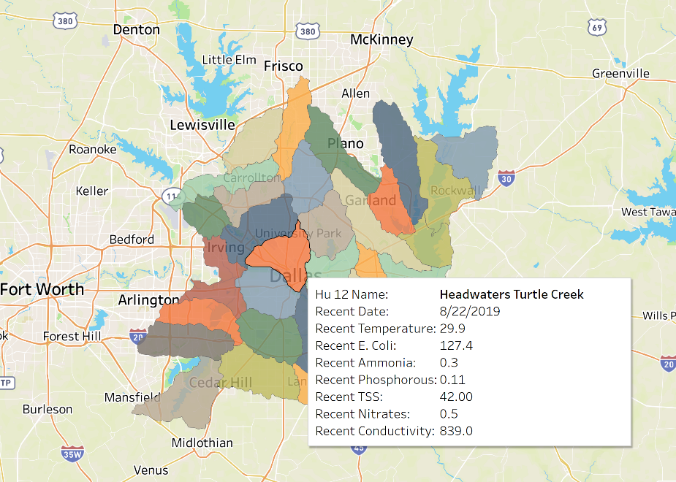
Watershed Drill Down – Contains a Bar Chart with summary statistics, a time series arranged by Sample ID, and an anomaly detection chart. This micro-level approach allows users to investigate one watershed at a time.

Watershed Comparison – Contains a time series, a table of Composite Scores, and a Box and Whiskers plot. This meso-level approach allows users to compare multiple watersheds at a time.

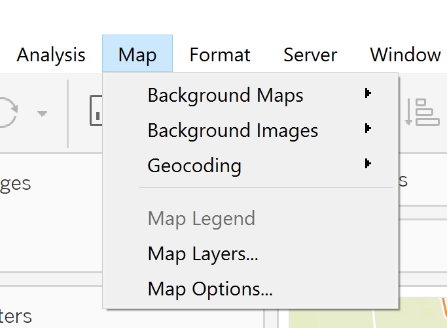
Each dashboard is a compilation of three to four *Sheets*. To declutter the file, sheets have been hidden.

However, users can unhide the sheets as desired. Hover on a dashboard item until the icon box (pictured left) appears in the top right corner. Select the second option from the top to route to the sheet. This allows users to temporarily view the sheet. The sheet will automatically hide again after clicking into another dashboard. To pin the sheet to the navigation pane, simply right click on the tab and select “Unhide Sheet.”

## GIS Map

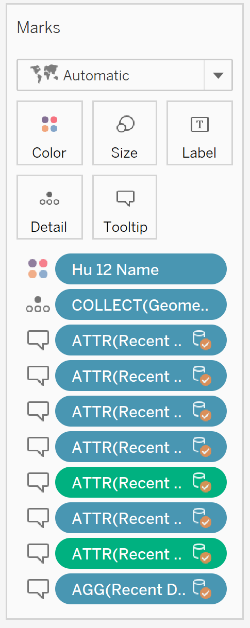
Data Source: GIS Shape File & WQ\_Clean[[1]](#footnote-2)

The Tooltip appears when hovering over a watershed. The values displayed are the most recent sample collections.



Alter the background map appearance by selecting the dropdown menu from “Map” in the toolbar. Select “Background Maps” to change the primary map layer. Users may select multiple options including a satellite image.

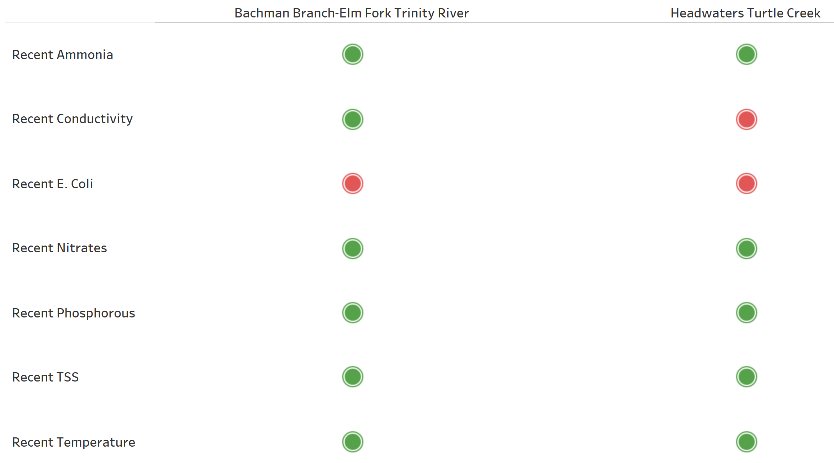
Select “Map Layers” to view additional options including the level of detail in the street names, region borders, water labels, and

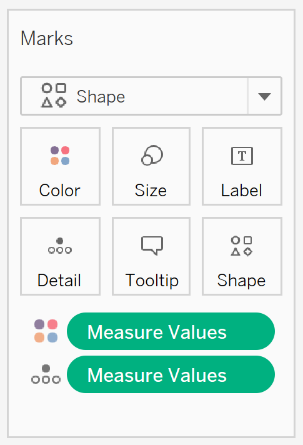


Use the “Marks” Card to further modify the appearance of the map. Select Color to change the color, outline, or opacity of the watershed shapes. Select Tooltip to change the values displayed when the user hovers his or her mouse on a watershed. New values can be dragged onto the Tooltip mark to add them to the existing list. Likewise, current values can be dragged off the Marks card to be removed from the display.

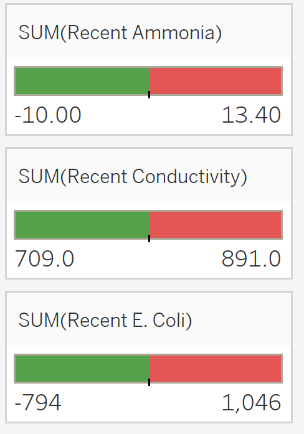
## Key Performance Indicator (KPI) Dots

Data Source: WQ\_Clean



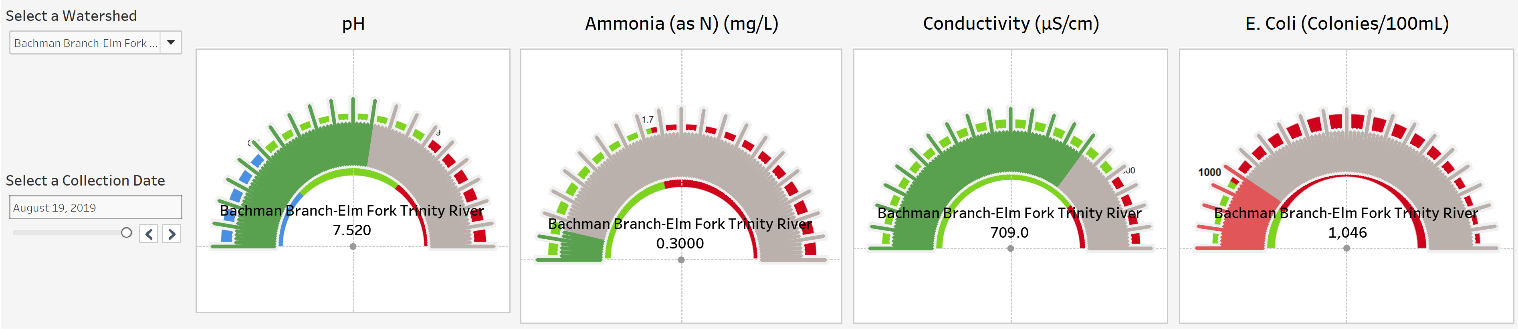


Use the “Marks” card to alter the shape and size of the KPI dots. Select *Size* to toggle the size of the dots. Select *Shape* to change the shape of the dots – Users can select from a range of shapes provided by Tableau, or upload custom shapes (instructions [here](https://www.tableau.com/drive/custom-shapes)).

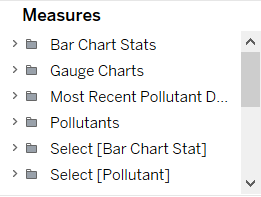


The color of each KPI dot is manually configured based on the values provided in the file “Parameter Acceptable Values.” To alter the color, hover over the top right corner of the designated color filter box. Click on the black arrow that appears and select “Edit Colors.” To change the colors only, click on the green or red boxes to the left and right. To update the Parameter Acceptable Value, change the *center* to the desired value.

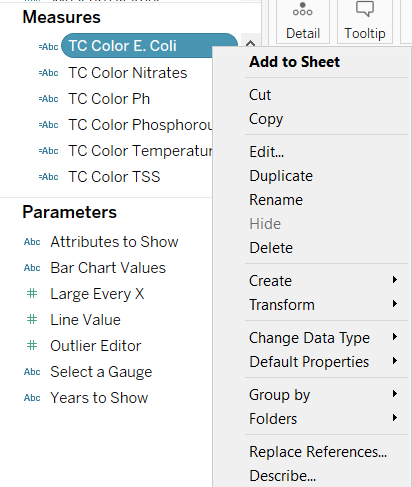
## Gauge Charts

Data Source: WQ\_Clean & Gauge Image

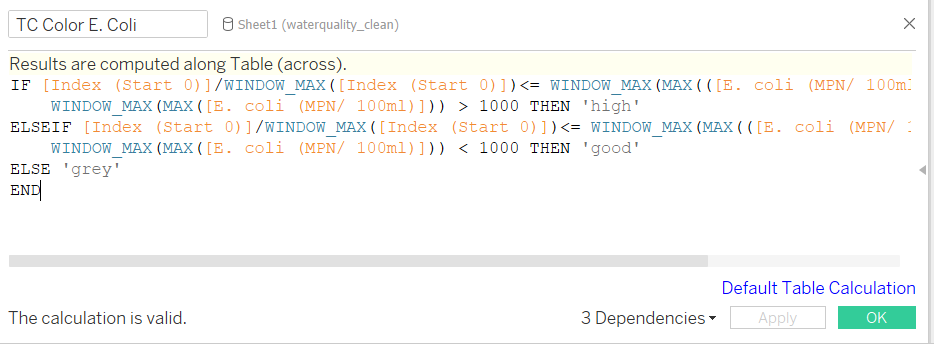
Click the arrows to browse through 10 years of water quality data. Gauges update in real time.

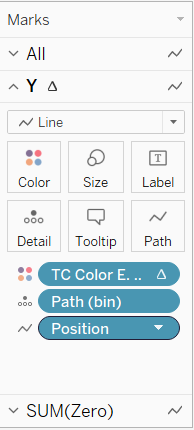


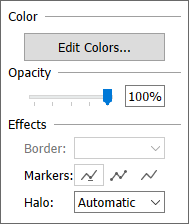
To change the acceptable parameter value for each pollutant click on “Gauge Charts” under Measures.



Right click on “TC Color” for the pollutant you want to change and select Edit (Pictured left - E. coli as an example).

The values inside the red boxes are the parameter limits provided in the “Parameter Acceptable Values” Excel file, but they can be changed it to any desirable number. Be careful to update both values located immediately before each “THEN” clause. Note that pH is a special case since there is an upper bound (9) and a lower bound (6.5). Thus, there are three values that need to be updated when altering the limits for pH: the upper bound, the lower bound, and the space in between.

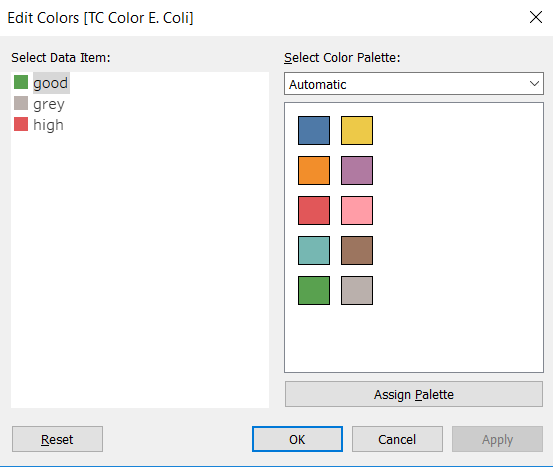
To edit the color of the gauges first click on the *Marks* card named “Y”



Then click on ‘Edit Colors’

Then click on the ‘Color’ editor here.

## Use this interface to manage the color of each gauge chart. There are multiple color palettes provided in Tableau. Choose the palette by selecting the drop-down menu from “Select Color Palette.”



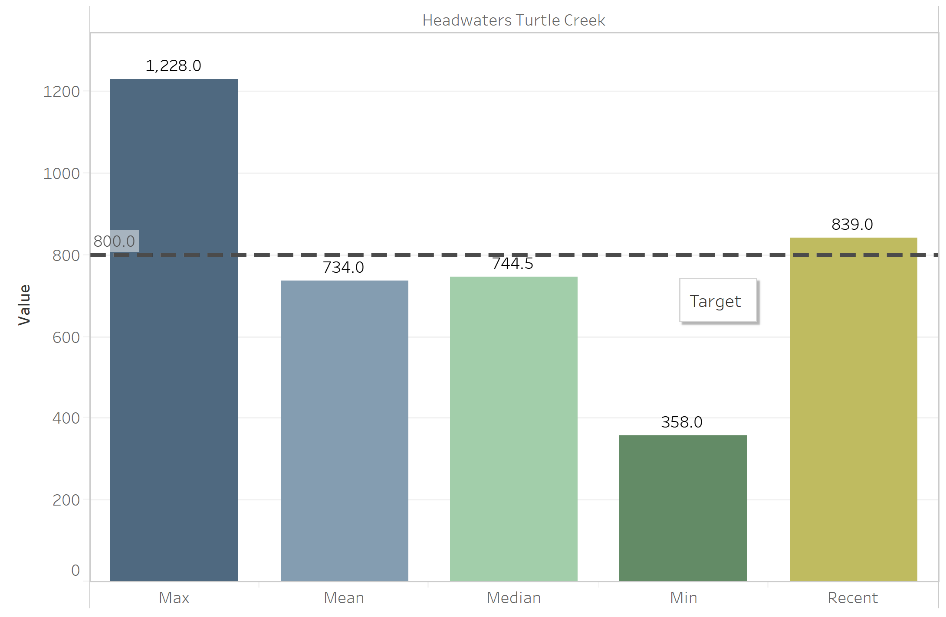
## The gauges are configured to turn green when the value meets the acceptable parameter value and red when the value exceeds the acceptable parameter value. Each gauge is configured with unique upper and lower bounds. Thus, the “grey” area corresponds to the neutral space in the chart. If a gauge chart has no grey space, then the value completely exceeds the upper bound. This should only occur with extreme outliers.

## 

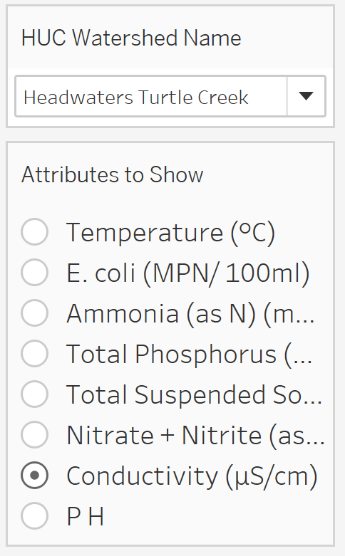
## Bar Chart

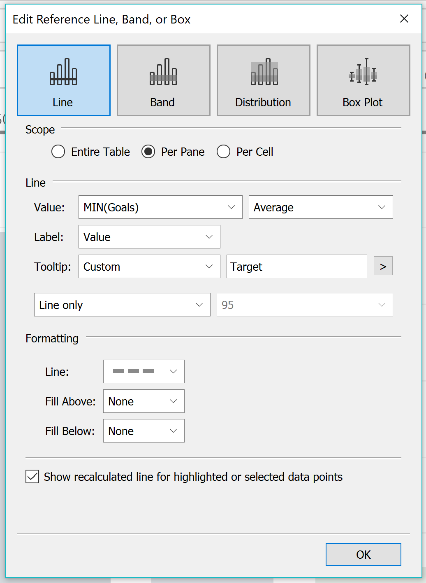
Data Source: WQ\_Clean

Grey Dashed Line indicates the Parameter Acceptable Value

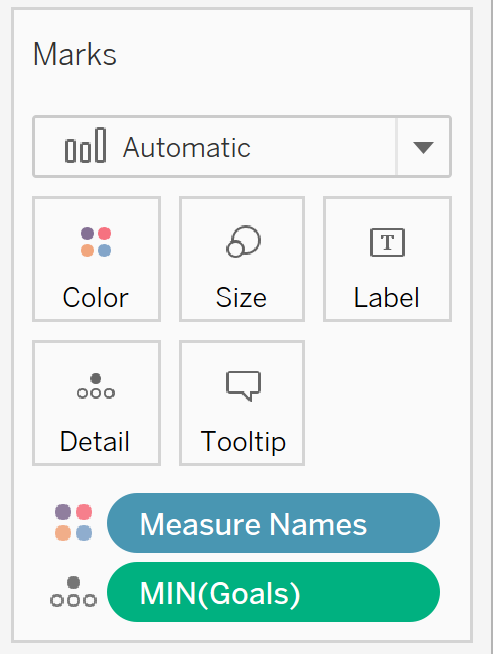


Use the filters to select the desired pollutant and watershed. Bar Chart values are manually configured calculations. Note a few key manipulations used to configure this chart:

1. The Max, Mean, Median, and Min are calculated using almost every available data point provided in the raw data file. The concurrent samples from identical sample sites are included. Computing these statistics in excel will yield nearly the same result with differences noticeable only in the thousandths.
2. The Recent value is the Maximum Most Recent Value. Often, multiple samples are collected from different sample sites on the same date. The visualization quickly becomes cumbersome if data from all sample sites are included. Instead, the Maximum Most Recent Value is displayed. This can be considered the worst-case scenario. Users are encouraged to use the Time Series chart to investigate further.



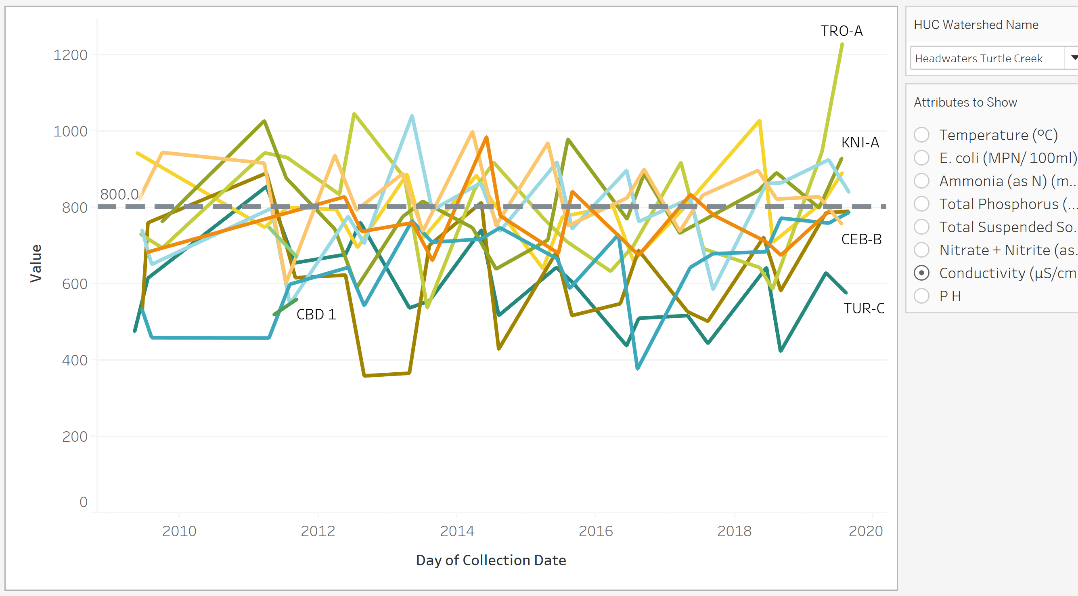
Alter the grey dashed reference line by right clicking on the line and selecting “Edit.” The dialogue box (pictured left) will appear. Configure the settings to adjust the *Label*, *Tooltip*, and line *Formatting*. Other settings should be left in the current state to preserve the intended functionality.



Use the “Marks” card to adjust the appearance of the bar chart. Adjust the *Size* to change the thickness of each bar. Use *Label* to toggle whether the values appear at the top of each bar. Use *Tooltip* to

## Time Series – Single Watershed

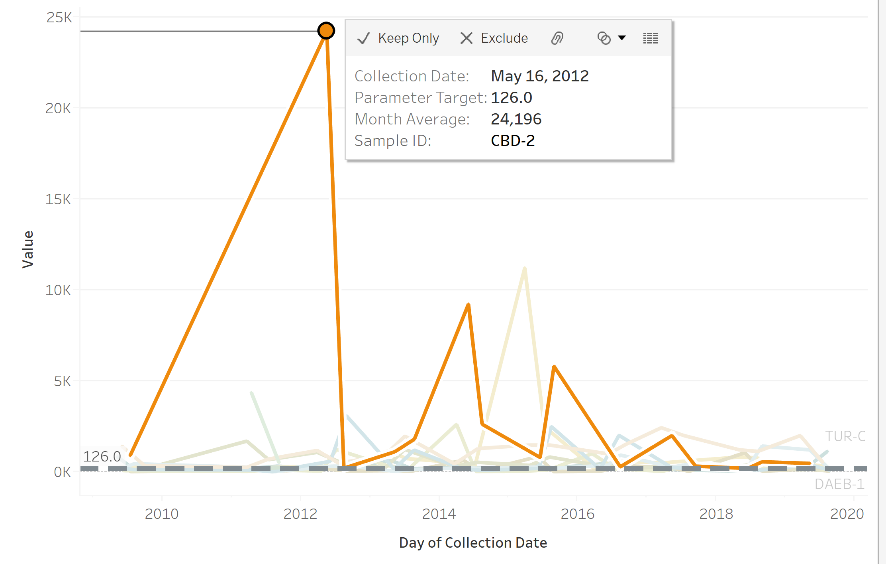
Data Source: WQ\_Clean



Select one Watershed at a time and Choose one Pollutant.

Grey Dashed Line indicates the Parameter Acceptable Value

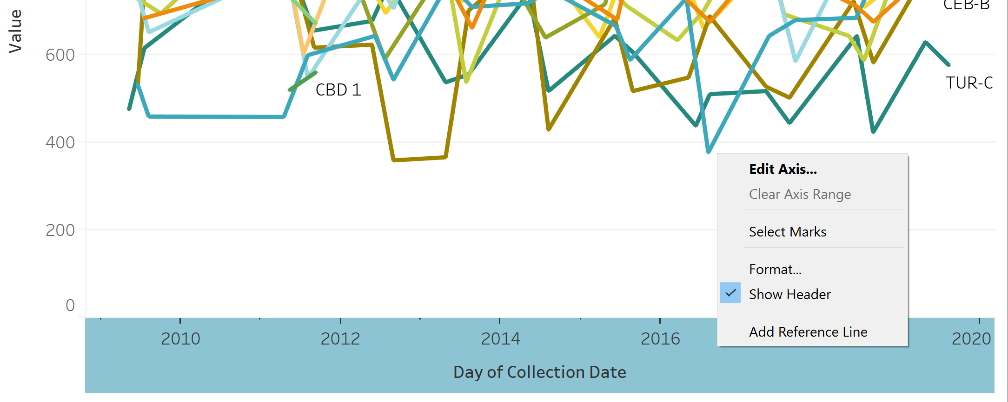
Although chaotic at first, the level of granularity in this chart provides immense value to the story behind the data.



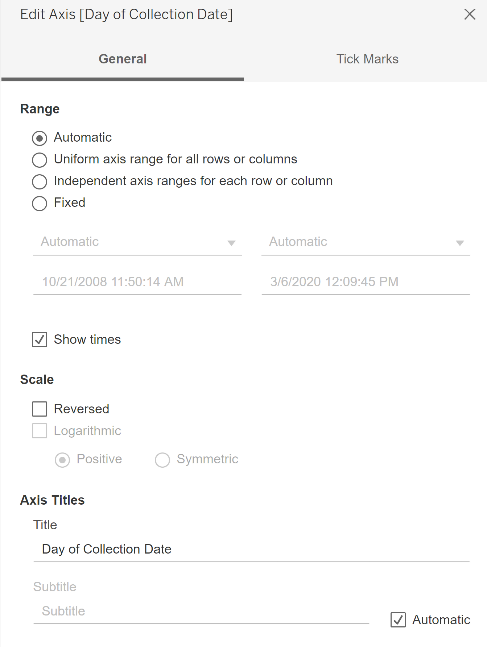
Clicking on a line will drown out the noise by fading the other lines in the background.

Click on a point to exclude specific lines or points. For example, the most egregious outliers in E. Coli can be excluded from the chart. After excluding values, use CTRL + Z to undo the changes and revert back to the original graph.

Dynamic Zooming

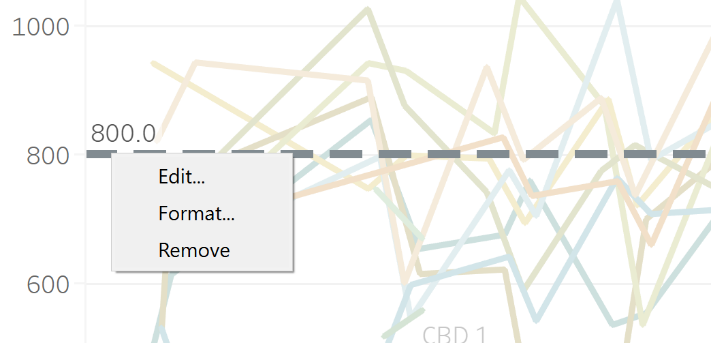
Hold down CTRL and use the scroll wheel on your mouse to zoom in and out of specific sections of the chart. Note that this will temporarily adjust the X and Y axis scaling.

When the user selects a new pollutant, the axes will no longer scale automatically. To fix this, simply right click on the X and Y axis. Select *Edit Axis.*

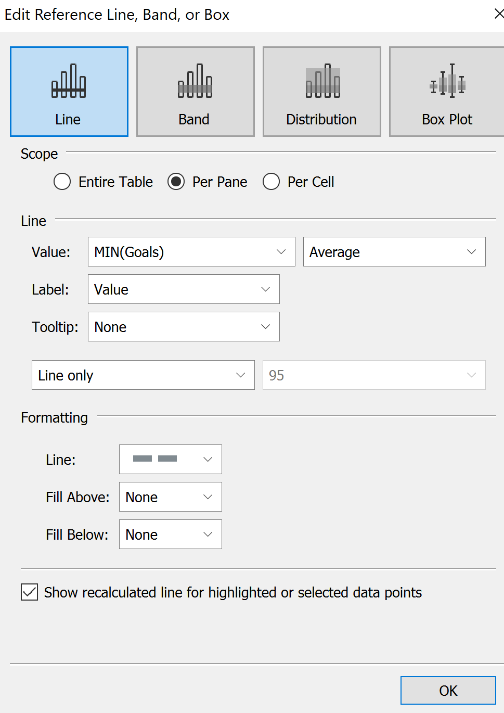


In the dialogue box that appears, select “Automatic” to update the axis so it will scale automatically upon choosing another pollutant. Do this for both the X and Y axes separately.

The dialogue box can also be used to set a fixed axis range, scale logarithmically, or change the axis titles.

Alter the Reference Line

Currently, the reference line displays the Acceptable Parameter Value. However, it may be useful to view the average to track how it changes after excluding outliers. Right click on the grey dashed line and click *Edit.*

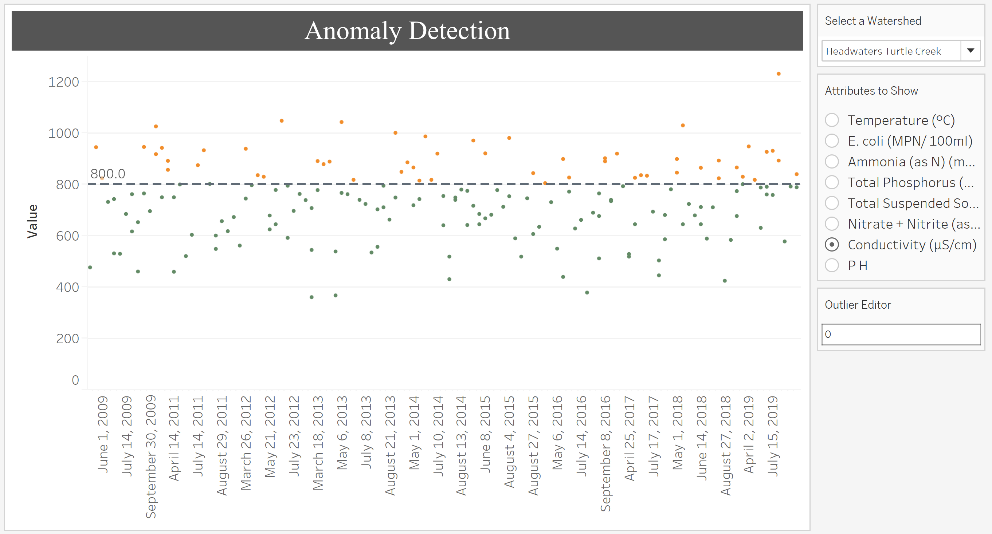


In the dialogue box that appears, changes the Value to “Measure Values.” MIN(Goals) is a parameter that contains the Acceptable Parameter Values.

The dialogue box can also be used to configure the Tooltip that appears when users hover the mouse over the reference line. Adjust the *Formatting* to change the color of the line, the dash thickness, or color the area above or below the line.

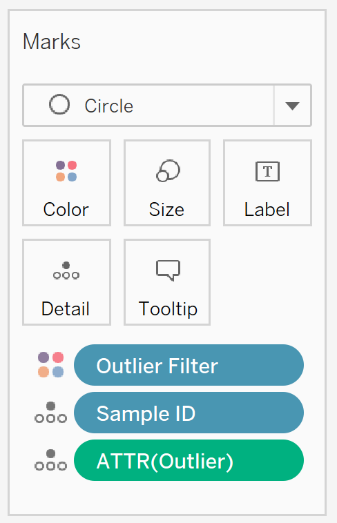
## Anomaly Detection

Data Source: WQ\_Clean



Add or subtract an integer value to the “Outlier Editor” to adjust the line dynamically

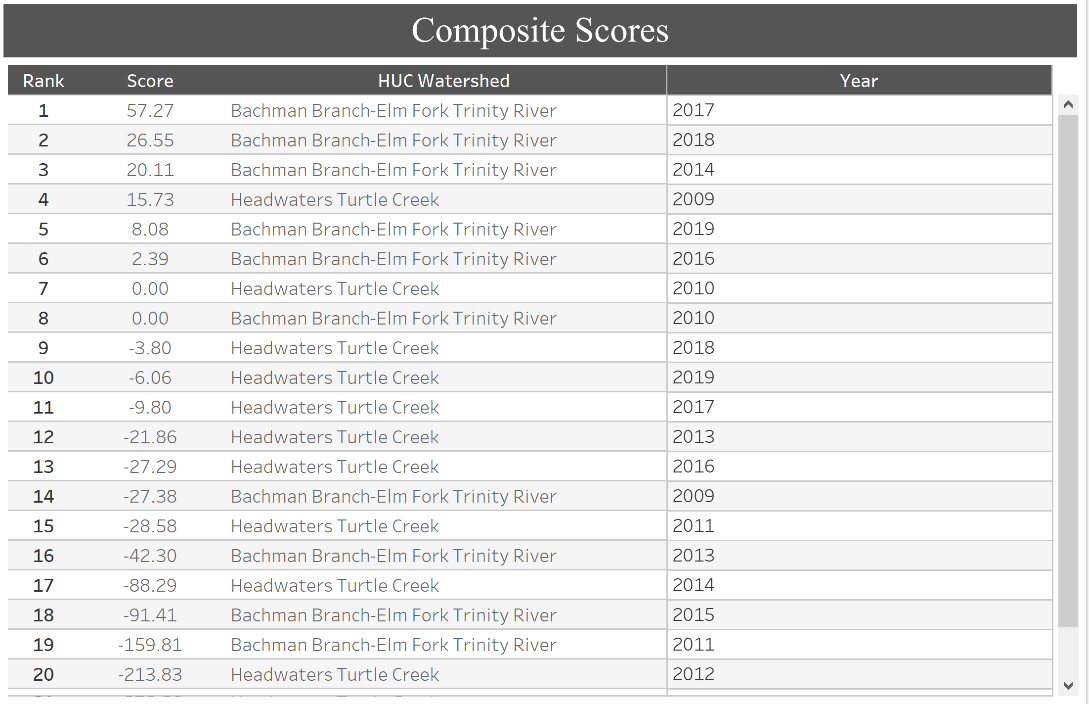
Dashed line shows the Acceptable Parameter Value. Points above the line are orange, points below the line are green.



Use the “Marks” Card to change the color, shape, and size of the data points. Select “Tooltip” to alter the message that appears when users hover their mouse on a specific data point.

## Composite Scores

Data Source: Ranked\_Scores [[2]](#footnote-3)



Alter Color Scheme:

Right Click on the title and select Edit to change the background color, font style, and font size of the title.

Right click on a column header and select Format to open the dialogue box to configure the headers.

Right click on a row and select *Format* to open the dialogue box to configure the rows. Click on the paint can and look for “Row Banding” at the bottom. This feature allows the user to pinpoint exactly which columns and rows should be colored.

1. .xlsx file created from R Code [↑](#footnote-ref-2)
2. Created in R File [↑](#footnote-ref-3)