Introduction:

The available dataset includes several years of water sensor readings from rivers and streams in theBoonsong Lekagul Wildlife Preserve. The samples were taken from ten different locations scattered throughout the area and contained values of several measures recorded over a period of 19 years.

The visuals created to answer the analysis questions based on the data are made using the visualization tool Altair on Jupyter Notebook. This report focuses on the trends, anomalies, missing data, collection frequency and unrealistic values found in the dataset.

Findings:

1. **Trends:**

   i) Clear seasonal trend with water temperatures starting low, peaking in June, July, August, and decreasing again across all years and locations.

   ii) Decrease in total hardness in 2011 and drastic increase from 2012 to 2014 in Boonsri, Kannika, Kohsoom, and Sakda.

   iii) Increase in aluminium from 2008 to 2009 followed by a substantial decrease by 2010 in all locations that aluminium was recorded.

2. **Anomaly:**

   Anomaly detected in iron levels in 2003 in certain locations

3. **Missing Data:**

   i) Ony 9% of the total measures were recorded in every year and location.

   ii) Locations Achara, Decha and Tansanee do not begin taking water readings till year 2009.

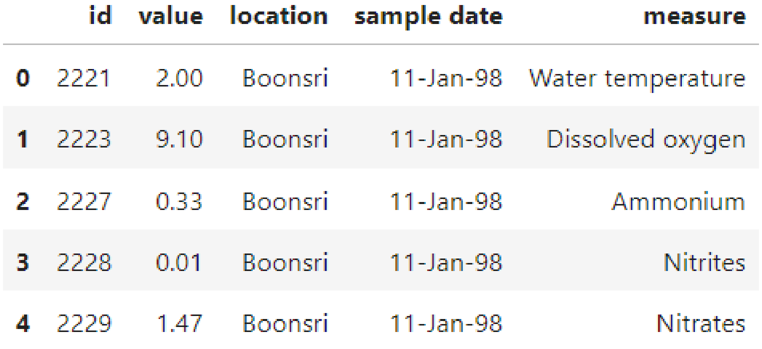
4. **Change in Collection Frequency**

   There is inconsistency in the frequency of records taken across all years and locations.

5. **Unrealistic Values:**

   Setting 0.1% contamination rate, outliers were recorded in the measures Total Dissolved Salts, Total Coliforms, Manganese, Aluminium, Iron, Copper, Zinc and Fecal Coliforms.

Before commencing the analysis, data preprocessing steps were executed, including loading the dataset "Boonsong Lekagul waterways readings.csv" into a Pandas DataFrame. The dataset, comprising water quality measurements, was examined for missing values, and an exploration revealed a wide range of measures.



	id	value	location	sample date	measure
0	2221	2.00	Boonsri	11-Jan-98	Water temperature
1	2223	9.10	Boonsri	11-Jan-98	Dissolved oxygen
2	2227	0.33	Boonsri	11-Jan-98	Ammonium
3	2228	0.01	Boonsri	11-Jan-98	Nitrites
4	2229	1.47	Boonsri	11-Jan-98	Nitrates

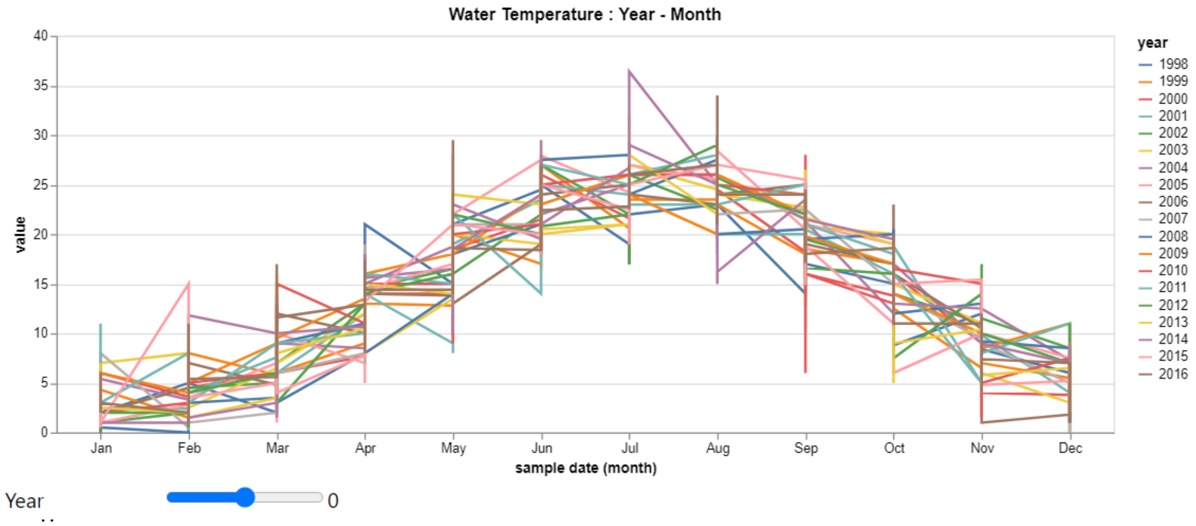
1. TRENDS/ANOMALIES

i. Trend:

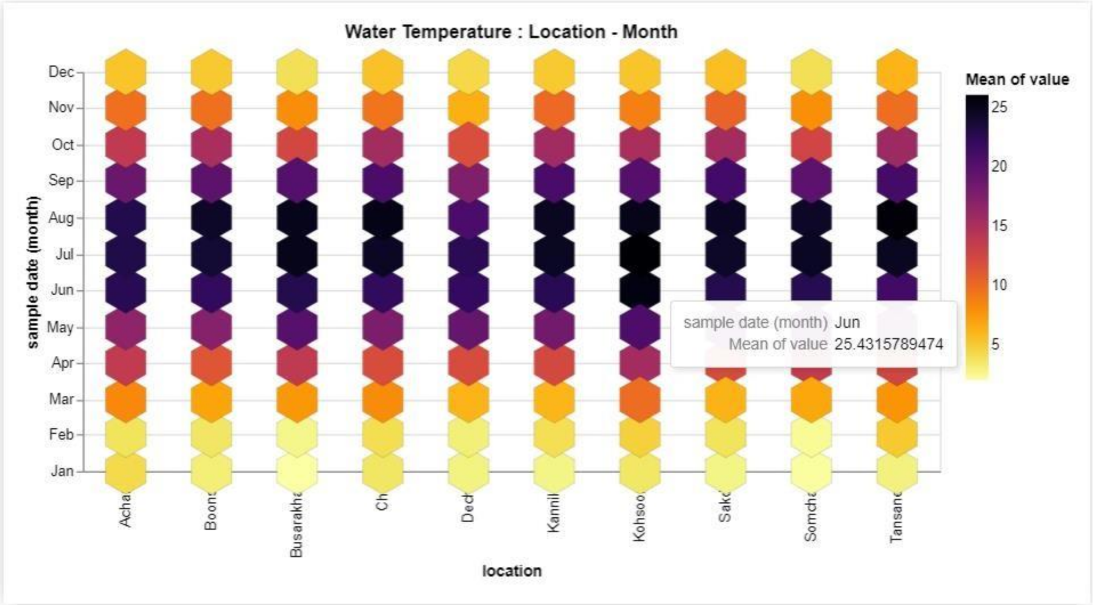
a) Water Temperature Trends

Finding: Clear seasonal trend with water temperatures starting low, peaking in June, July, August, and decreasing again across all years and locations.

Visualization 1: Month-wise temperature in different years
Line chart with months on x-axis, showing yearly patterns. Slider allows dynamic year selection. Each line corresponds to a specific year.



Visualization 2: Location - Month Water Temperature (Heatmap)
Hexagonal heatmap representing mean temperatures in every location and month, with warmer colours denoting higher temperatures.



- How can the finding be seen from the visualizations?

Both visualizations illustrate that the water temperature consistently starts low, peaks in June, July, and August, and then decreases towards the end of the year. The line chart captures the temporal trend, while the heatmap emphasizes the intensity of temperature variations on a monthly basis.

- Any advanced Altair visualization features used, such as multi-layer, chart concatenation, and interaction?

The line chart includes a slider for selecting specific years (interaction), allowing a closer examination of the trend for individual years. The heatmap employs a hexagon shape for data points, adding a distinctive visual element.

- Why is the chart type most appropriate for the analysis?

Line charts are the most effective type of charts for displaying temporal trends, making it suitable to showcase how water temperature changes over time. It helps in illustrating the overall trend in water temperature, making it easier to identify patterns and peaks. Heatmaps are an effective way of presenting variations in the data across multiple dimensions. Here, to compare the water temperature in different locations and years, the colour encoding enhances the magnitude helping to quickly identify periods of high or low temperatures.

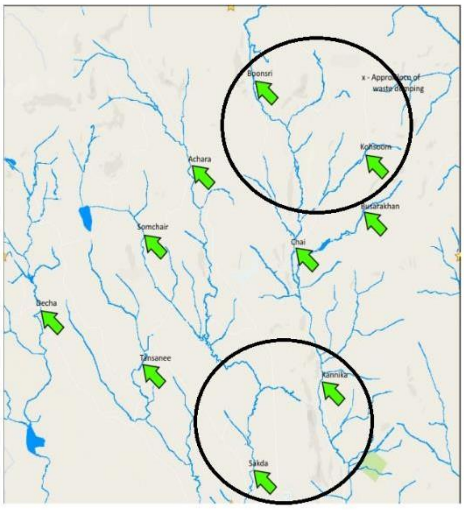
b) Top 5 Water Quality Indicators Trend: Location and Year

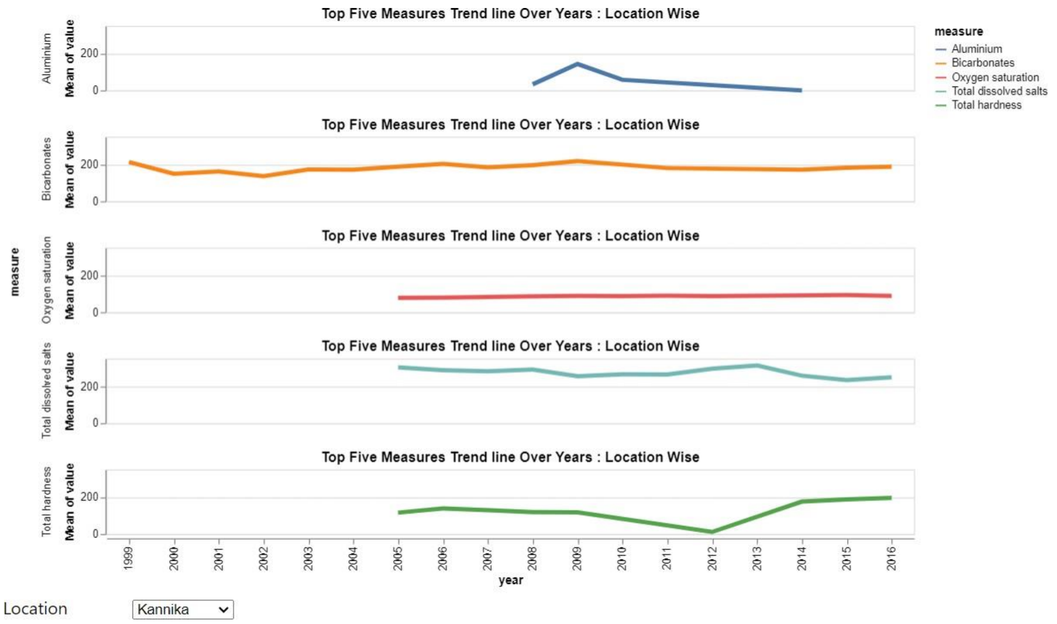
All measures were first grouped based on their mean values in the dataset and was then sorted to come up with the top 5 indicators. Aluminium, Bicarbonates, Oxygen Saturation, Total Hardness and Total Dissolved Salts were the measures with the highest average in the dataset. To look for their trends in the water year wise and location wise, a trend line graph was created with each facet representing a different measure.

Finding 1: Decrease in total hardness in 2011 and drastic increase from 2012 to 2014 in Boonsri, Kannika, Kohsoom, and Sakda.

Finding 2: Increase in aluminium from 2008 to 2009 followed by a substantial decrease by 2010 in all locations that aluminium was recorded.

All other measures (bicarbonates, oxygen saturation, total dissolved salts) show almost the same mean value in all years and locations.





Location Kannika

- How can the finding be seen from the visualizations?

The trends are clearly seen on the graphs at one glance. The increase of aluminium level from 2008 to 2009 and a decrease from 2009 to 2010 can be seen on the line showing aluminium trend, and total hardness decreasing from 2010 to 2011 and the increasing after 2012 can be seen in the locations Boonsri, Kohsoom, Kannika and Sakda. The location selection can be done on the dropdown box.

- Any advanced Altair visualization features used, such as multi-layer, chart concatenation, and interaction?

Dropdown for location selection allows interactivity and focused exploration on each location. Facet layout makes the comparison of trends across measures compact yet detailed.

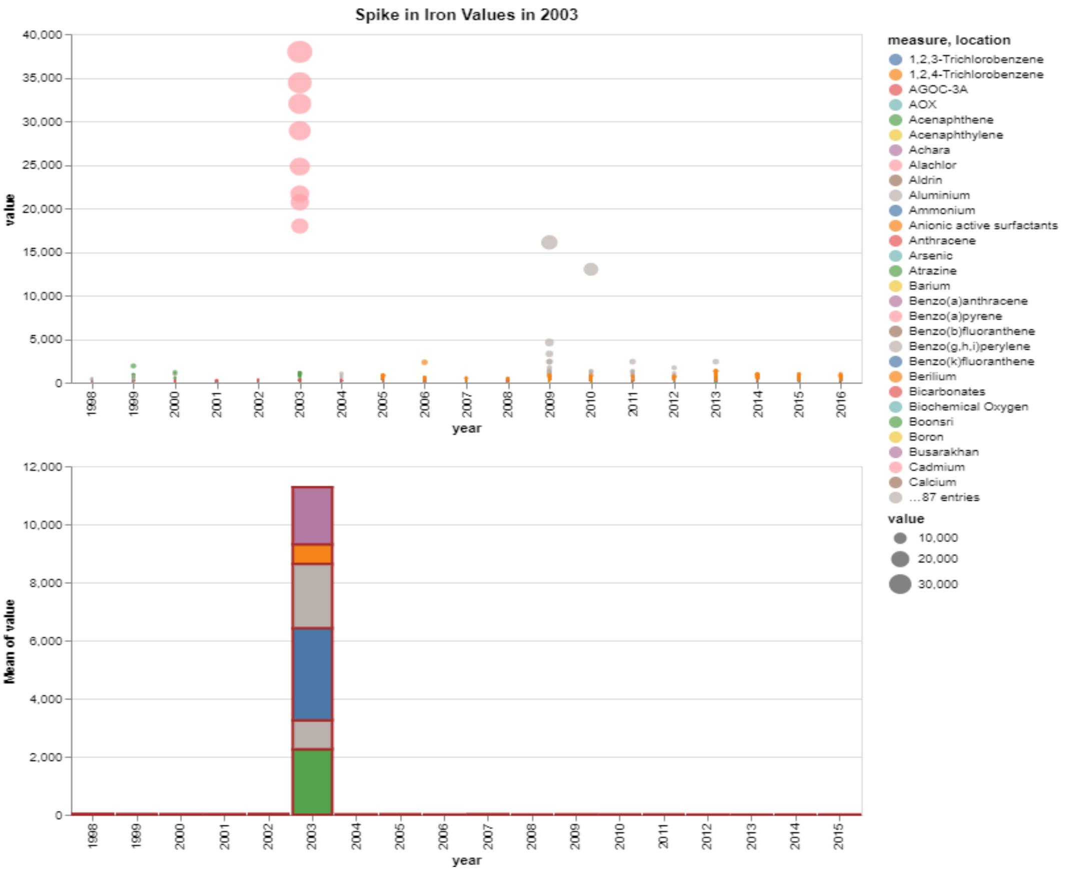
- Why is the chart type most appropriate for the analysis?

A line chart is a straightforward way to ensure that the graph is interpretable without complexity. The lines provide a clear path making it simple to identify upward or downward trends, peaks and variations. The faceting of each measure makes it easy to differentiate between them but also at the same time make comparisons easier.

ii. <u>Anomaly:</u>

Finding: Anomaly detected in iron levels in 2003 in certain locations

- The scatter plot (Chart 1) using the entire dataset reveals **high iron values in the year 2003.**
- The bar chart (Chart 2) focused on iron as a measure confirms a substantial increase in iron levels in 2003 **across multiple locations**.



Very evidently from the two charts, it was seen that iron values stood out in 2003 compared to the other measures in all the years combined. The bar chart was then created to check what constituted the high levels of oxygen. It was seen that on August 15th 2003, the level of iron content in Busarakhan, Chai, Kannika, Kohsoom, Sakda and Somchair was extremely high compared to those onother sample dates. While in all other years the mean of iron content was between 0 and 2, in 2003 these locations had mean between 500 and 4000. This is clearly an anomaly.

- How can the finding be seen from the visualizations?

Scatter plot:

The scatter plot shows a concentration of larger circles in the year 2003, indicating higher iron values compared to other years. The spike in iron values during 2003 is visually distinct, signaling an anomaly.

Bar chart:

The bar chart specifically focused on iron levels confirms a noticeable increase in mean iron content across multiple locations in the year 2003. The brown bars for the year 2003 stand out, emphasizing the anomaly in iron concentrations.

- Any advanced Altair visualization features used, such as multi-layer, chart concatenation, and interaction?

The scatter plot includes a tooltip that provides additional information about each data point, specifically the measure. This feature enhances engagement and allows for detailed examination of individual data points. The bar chart is combined with the scatter plot using the Altair composition operator **&** to create a single display with both charts. This allows a comprehensive view of iron levels across different locations and years in one visualization.

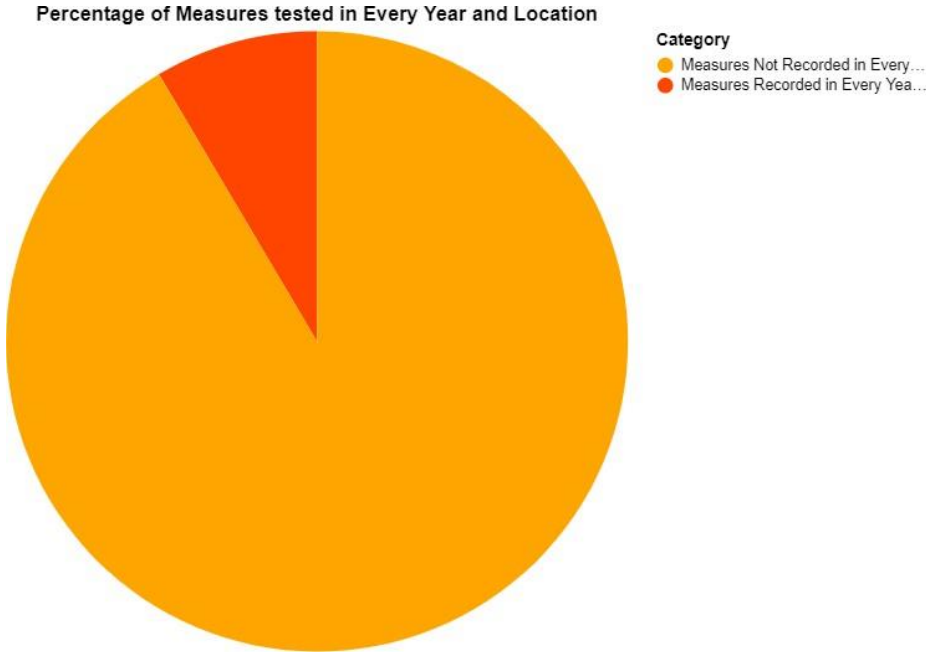
- Why is the chart type most appropriate for the analysis?

The scatter chart to represent individual data points makes it effective for highlighting specific situations, like the anomaly here. The anomalous points in 2003 can be visually distinct from other data points. The bar graph is the most suitable chart to show the iron levels across different locations in the year 2003 because the comparison is being done on aggregated values and the abnormal increase in the level of iron can stand out in the chart compared to other years and locations.

2. DATA QUALITY AND UNCERTAIN ISSUES

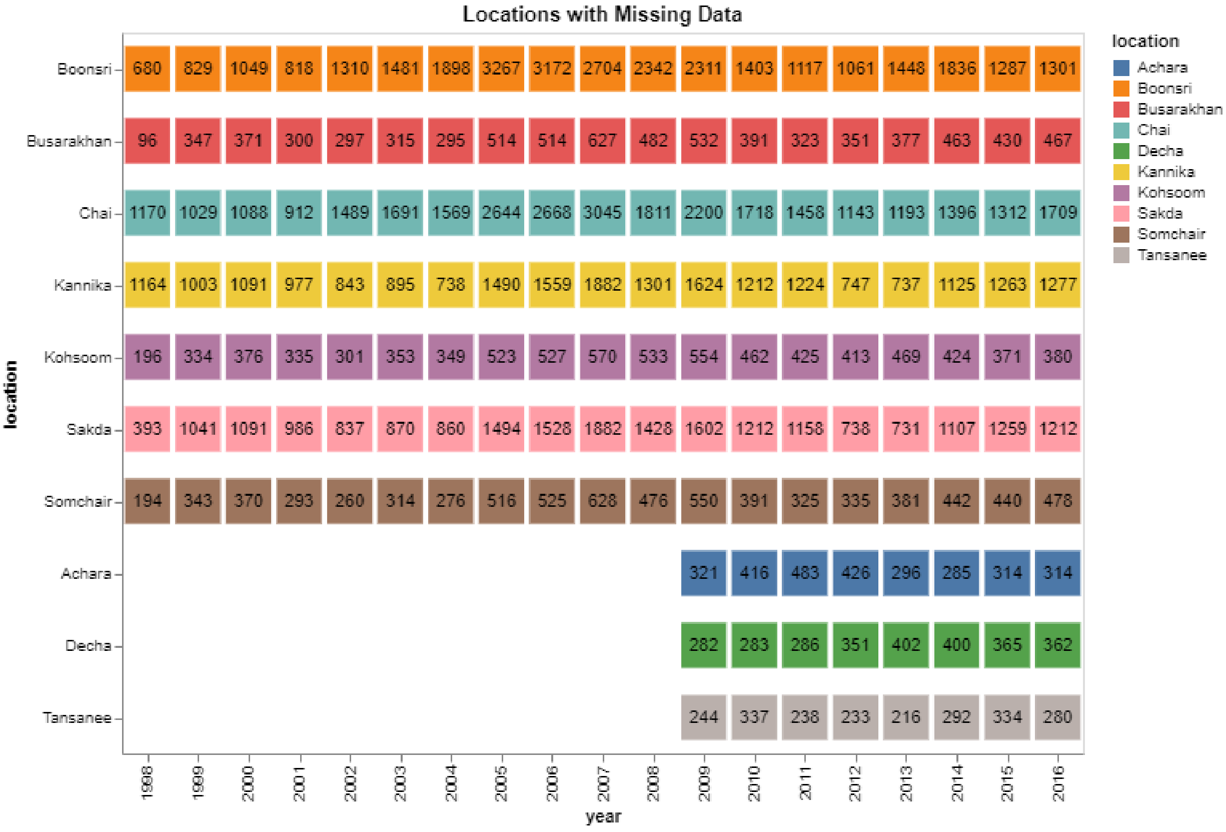
i. Missing Data

From 1998 – 2016, a total of 106 measures were tested in 10 different locations. Ideally, for the best analysis it is crucial that all 106 measures are tested in every location and year in the dataset. This ensures that the analysis captures all variations and patterns that may exist in the dataset.
So, to check the consistency of data recording across different locations and years, a pie chart was created that showed the percentage of the total measures that satisfied this condition.



From the chart, it is very clear that **only a very small number of measures were actually tested in every location and year. To be precise, using the tooltip we can see that only 9, that is, only about 8.4% of the total measures were tested in all the years and locations.**

Also, a heatmap was created to check the number of records that were taken in each location and year.



This chart summarizes the variation in the number of records taken each year in every location. The text helps identify the exact number of records. Most importantly, it is evident from this graph that **the three locations Achara, Decha and Tansanee do not begin testing until 2009**. This is important because it could leave out important information about the waterways.

- How can the finding be seen from the visualizations?

From the pie chart it is clear that only a small part of the total number of measures has been tested in every year and location. This accounts for a large portion of the missing data. Furthermore, the heatmap shows how three of the locations had not started taking water readings from 2009. To accurately analyze the data, it needs to be consistent and reliable.

- Any advanced Altair visualization features used, such as multi-layer, chart concatenation, and interaction?

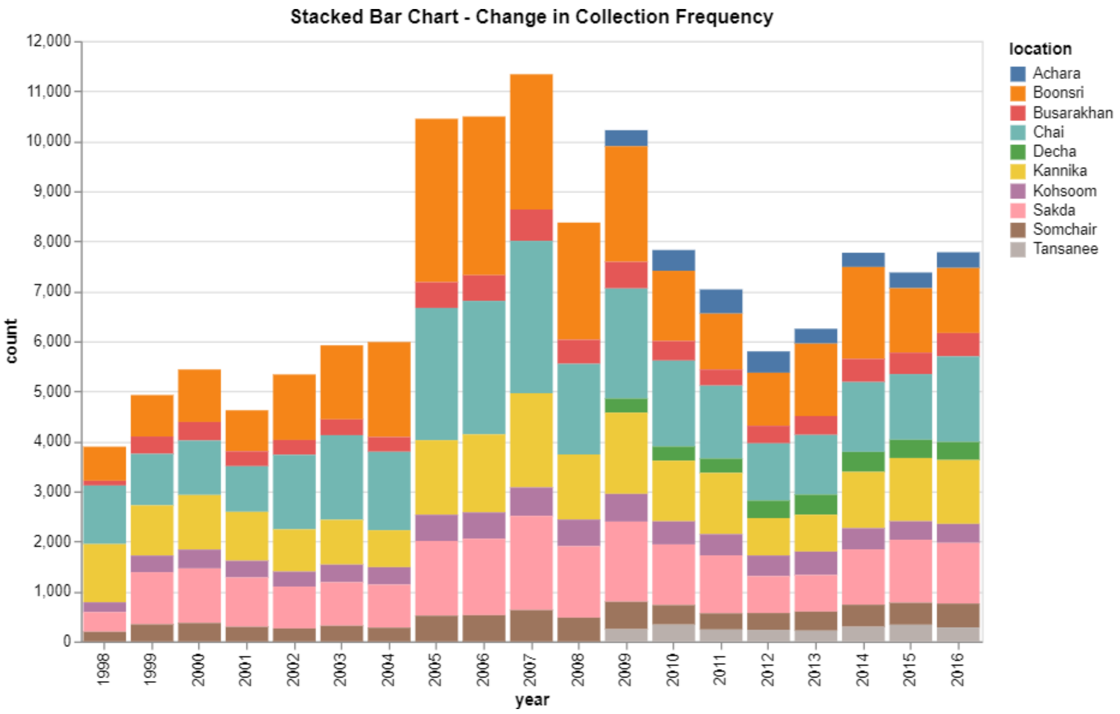
Pie chart uses the tooltip feature. It shows information about the number and percentage of measures. Heatmap uses the text feature where count of records in each year and location is displayed on each square.

- Why is the chart type most appropriate for the analysis?

Pie chart was used to visualize the percentage of measures that were tested in every location and year because the size of the wedges makes it easier to compare the proportions of different categories. The usage of heatmap to visualize counts of records in each location year wise makes it easy to spot missing data. Colour channel is used to differentiate between locations.

ii. Change in Collection Frequency

To find the change in frequency of data collection in all locations over years, a stacked bar chart is created with each bar representing a year and each stack in a bar representing a particular location.



This chart depicts the total observations made in different locations in the years 1998-2016. Using the colours in the stacks, it can be seen that out of the ten locations only seven have observed measures from 1998-2016. Rest of the locations have observations only from 2009. The location with the largest total measure count is Boonsri. **In all the locations the greatest number of observations were made in the period 2005-2009. This is because in all the locations, in the year 2005 there is almost double the number of records taken as compared to 2004**. **However, the frequency of collection of data went down again after 2009.**

- How can the finding be seen from the visualizations?

From the visualization, **there is a clear uneven distribution of stack heights which shows the inconsistency in the measure recording**. In all locations, initially the frequency of sampling was less but it rose in the years after 2004 but decreased again after 2009. Inconsistent frequency of sample recordings can influence the interpretation of the true effects of the measure.

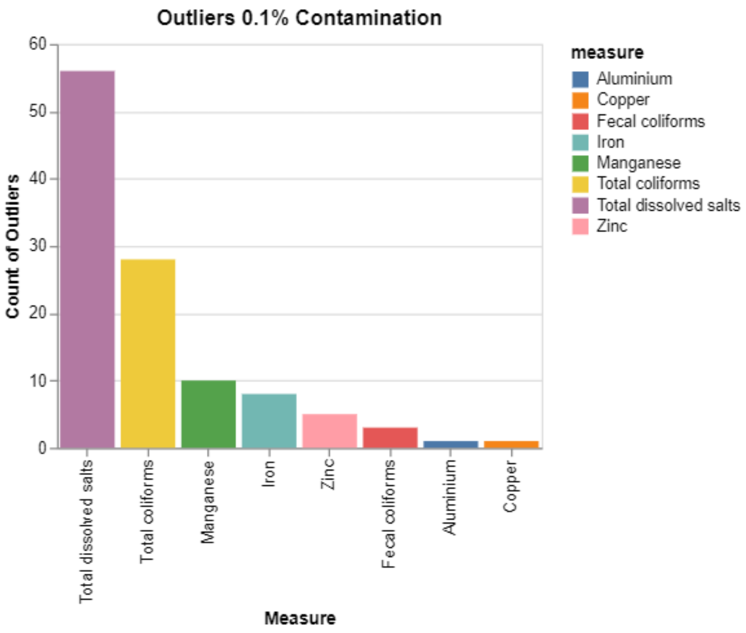
- Why is the chart type most appropriate for the analysis?

The temporal trends are clearly presented on the x-axis enabling an easy assessment of how the frequency of data collection evolved over the years. The division of the chart into stacks for each location ensures detailed analysis, distinguishing between locations. Also, the use of colours in the stacks increases visual differentiation, making it easy to understand each location’s collection history.

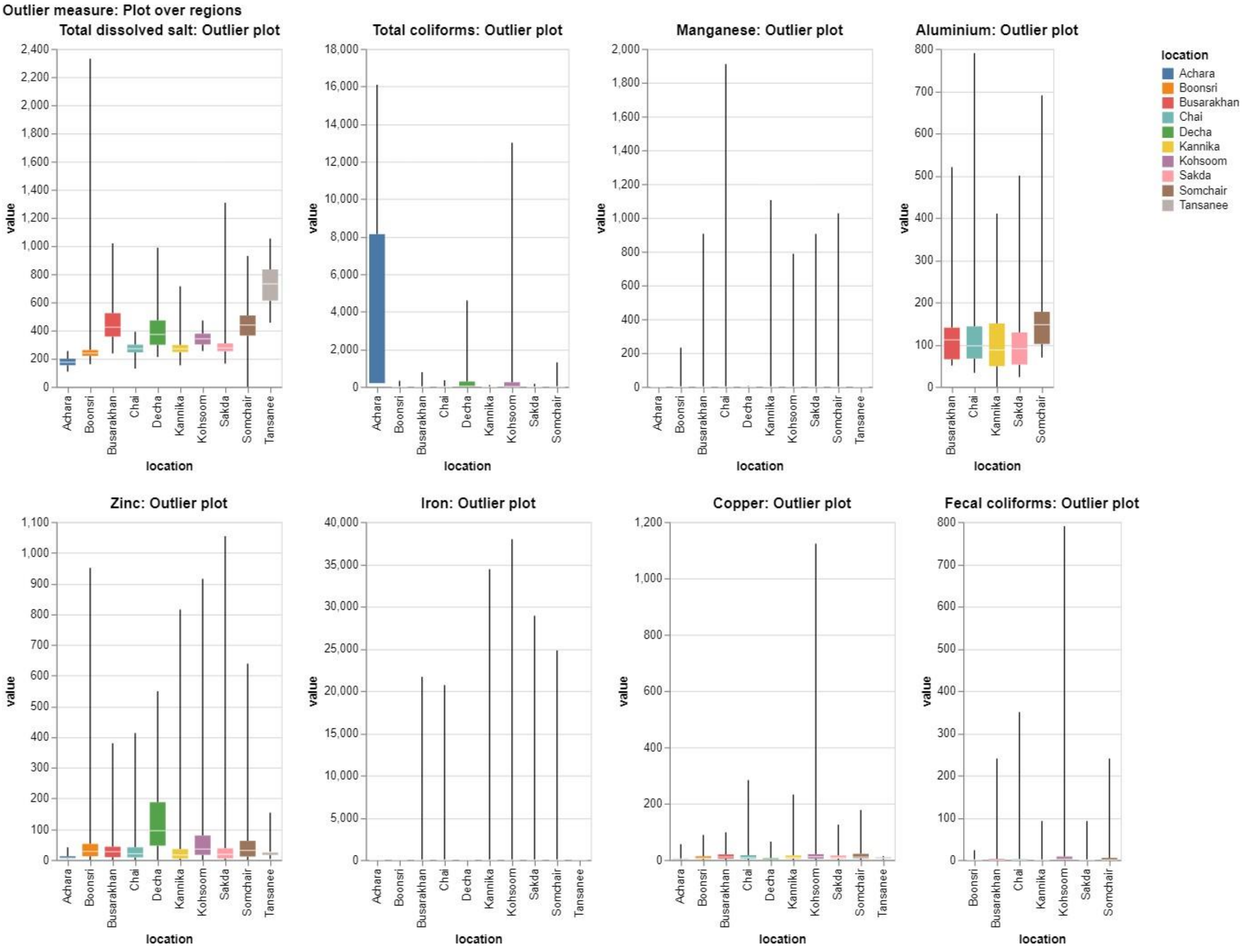
iii. Unrealistic Values (Outliers)

In the process to detect and visualize the outliers, a custom function, **detect_and_visualize_outliers**, was applied to the dataset. This function utilized the **Isolation Forest algorithm with a contamination rate set at 0.1%** to identify outliers within the dataset. The dataset was initially aggregated based on the measure, computing the mean value for each group. Subsequently, the Isolation Forest algorithm was employed to flag outliers in the dataset. The resulting count of outliers for each measure was then visualized using a bar chart. The x-axis of the chart represents different measures, while the y-axis denotes the count of outliers. The colour distinction in the chart allows for clear identification of outlier counts for each measure.

The order of measures with the highest count of error observations is as follows: **Total dissolved salts (56), Total Coliforms (28), Manganese (10), Iron (8), Zinc (5), Fecal Coliforms (3), Aluminium (1), and Copper (1). Notably, the measure with the most error observations was Total dissolved salts, highlighting its sensitivity to outliers when a contamination rate of 0.1% is considered.**



Furthermore, for each of these measures, boxplots were generated to visually inspect the distribution of error observations in each location. Before creating the chart, different data frames are created for each measure that showed outliers. The resulting chart then provided a comprehensive view of how each of these measures' data is distributed in each location.



Total Dissolved Salts: Outliers in the observations for Total Dissolved Salts are evident in the location Boonsri, where an outlier exceeds the typical range significantly. Sakda, Busarkhan, and Decha also exhibit variations in their observations, deviating from the expected range.

Total Coliforms: Noteworthy deviations in observed values are observed in Achara and Kohsoom for Total Coliforms, showing unrealistic values. Decha and Somchair also display irregular observations compared to the normal range.

Manganese: With the exception of Achara, Decha, and Tansanee, all locations exhibit outliers in their Manganese observations, indicating values outside the usual range.

Aluminium: All locations that aluminium was recorded in show notable deviation from the normal range.

Zinc: Apart from Achara and Tansanee, all other regions display some outlier observations for zinc, deviating significantly from the expected normal range.

Iron: Iron observations show high errors in all locations except Achara, Boonsri, Decha, and Tansanee, where the values remain within the expected range.

Copper: Kohsoom stands out with considerable deviations from the typical values in certain observations for copper suggesting a notable change from the norm.

Fecal Coliforms: All locations show outliers in Fecal Coliforms, with Kohsoom showing the highest difference from the expected range.

- How can the finding be seen from the visualizations?

The visualizations provide a comprehensive understanding in identifying and analyzing unrealistic values or outliers. **The bar chart offers a quick overview, highlighting the measures with the highest counts of error observations. Total Dissolved Salts stands out as the most sensitive to outliers, with a substantial count of 56,** emphasizing the need for careful consideration in handling this specific measure. **The subsequent boxplots offer a more detailed exploration, looking at the distribution of outlier observations within each measure and across various locations.**

- Any advanced Altair visualization features used, such as multi-layer, chart concatenation, and interaction?

The use of horizontally concatenated multi-layered boxplots including the use of '|', provides an advanced visualization method. This efficiently offers insights into distributions of outliers across measures and locations.

- Why is the chart type most appropriate for the analysis?

The bar chart with measures sorted by count enhances readability and allows for quick identification of the most significant outliers. On the other hand, multi-layered boxplots are ideal for revealing the spread and concentration of outliers within each measure across various locations.

CONCLUSION

The data we have given us a glimpse into whether the chemicals released in the wildlife preserve could be harmful or raise concerns if a non-harmful chemical is released rapidly. However, with some missing data and limited knowledge about how each chemical affects the environment, we're left to observe the trends within the dataset. If we had more info, like detailed geography or consistent recordings for each chemical across locations and time, we could better understand the situation in the Boonsong Lekagul Wildlife Preserve.