CST4060 REPORT 2

From the VAST Challenge 2018 mini challenge 2, visualisations have been made on ALTAIR to find insights in the provided datasets and the following are the results.

1. Data quality and Uncertain issues

We visualised average values of data with readings of values over 0.5 using the code below.

```
# Filter the data to include only those measures with average 'value' > 0.5
filtered_measures = data.groupby('measure')['value'].mean().reset_index()
filtered_measures = filtered_measures[filtered_measures['value'] > 0.5]
filtered_df = data[data['measure'].isin(filtered_measures['measure'])]
```

i. Missing data and change in collection frequency.

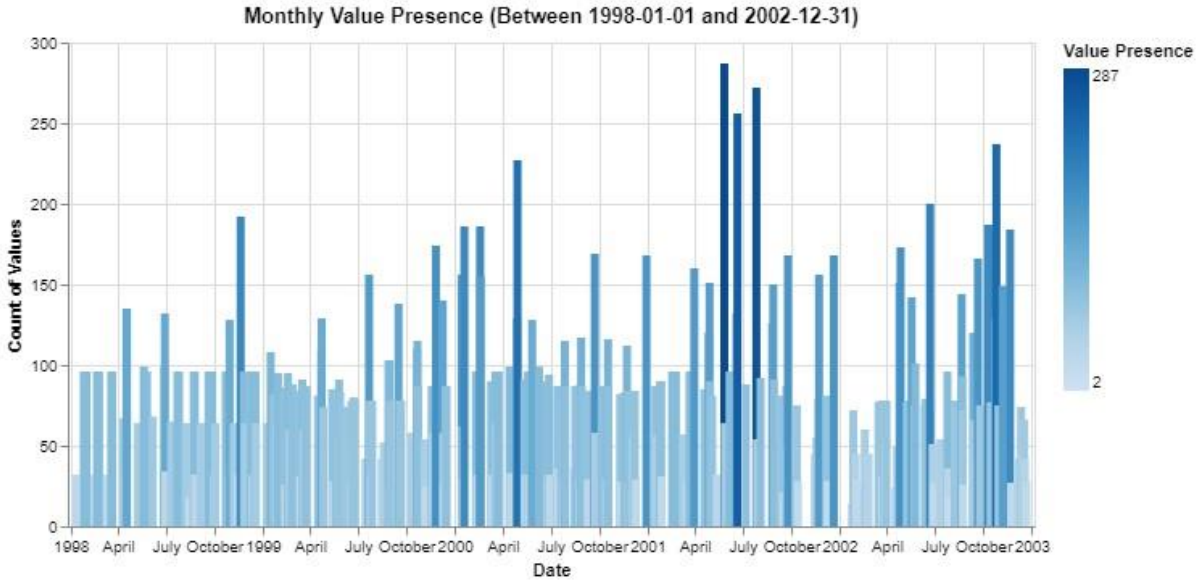
Findings: Missing data has been categorised in three ways:

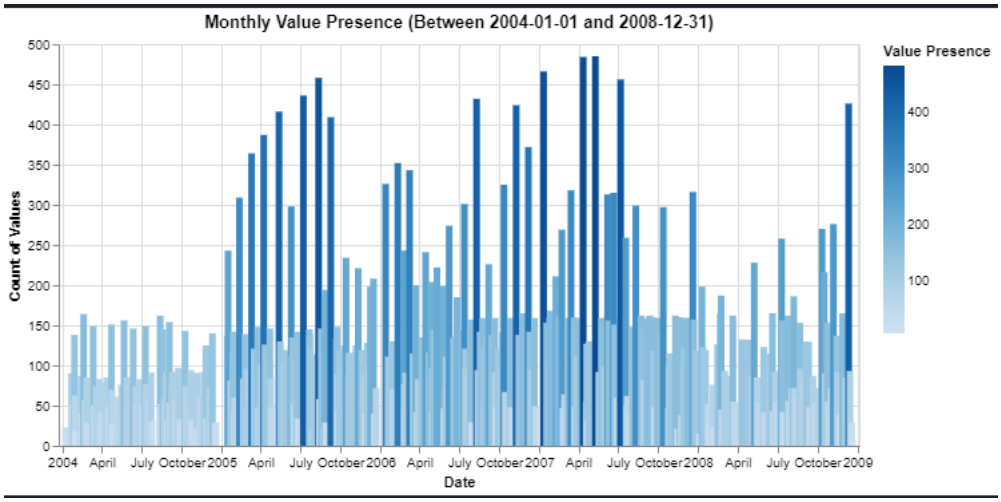
a. Null values: From the panda command data.info() , no null values exist in the datasets.

```
In [4]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 136824 entries, 0 to 136823
Data columns (total 5 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   id           136824 non-null  int64
 1   value        136824 non-null  float64
 2   location     136824 non-null  object
 3   sample date  136824 non-null  object
 4   measure      136824 non-null  object
dtypes: float64(1), int64(1), object(3)
memory usage: 5.2+ MB
```

b. Zeroes and dates with no values





From the two bar charts above, we can see high values of measures between 2004 – 2008.

Supporting argument was done using excel filter and functions. Zeroes occurred 9700 times in the dataset which is 7.09% of the available datasets. We have decided to leave these zeroes as it will not greatly affect comparison with other readings and that it is possible to get zero in some readings as we had readings between 0 and 1 appearing a lot of times.

Also filtering on excel, we discovered that Achara, Decha and Tansanee had data beginning only from 2009 up until 2016 but the rest locations had reading from 1998 – 2016.

ii. Unrealistic values

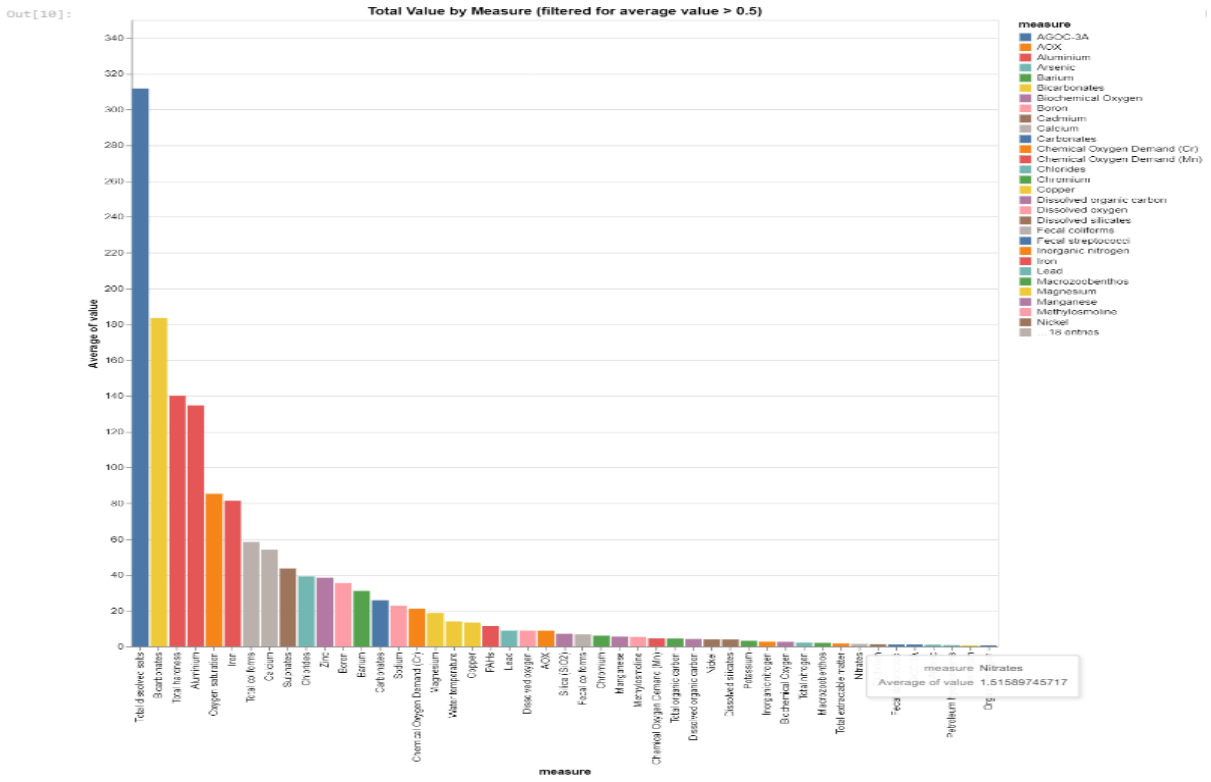


Figure 1: Bar chart showing average values of all chemicals

From the bar chart above, the bars have been sorted in descending order using the code below.

```
x=alt.X('measure:N', sort='-y')
```

A scale of 20 units was used used above, so values between 0 and 20 might be hard to read. This drawback has been taken care of with the use of tooltip code shown below.

```
tooltip=['measure:N', 'average(value):Q']
```

In the bar chart, the mouse was placed on Nitrates bar, and it showed the measure (Nitrates) and Average of the value (1.51689745717).

NB : The bar chart used here is just showing the average values of each chemical on an arbitrary scale and not comparing which is high or not as different measures can have different units. The chart shows the value of all chemicals especially when the mouse pointer is hovered on the bar.

Findings

The scatter plot below shows some unrealistic values for iron, total coliforms and manganese. From figure 1 above, average values of these measures can be seen in the table below.

CHEMICALS	AVERAGE VALUES
IRON	81.34
TOTAL COLIFORMS	58.44
MANGANESE	5.53

Table 1; Average values of iron, total coliforms and manganese

On August 2003, some unrealistic values of iron were recorded, this could be due to faulty equipment reading across different locations on that day. For instance, Kohsoom, Kannika and Sakda had readings of 37959.28, 34413.96 and 28901.59 respectively which is higher than the 81.34 average.

Unrealistic values for total coliforms, especially in Achara (value = 16090 , date ; Jan 15, 2009) and Kohsoom (value = 13000, date : October 20,2010) as shown in the scatter plots can be checked by placing the mouse on the plots. These values were equally higher than the average values.

Also, with average value of Manganese at 5.53, unusual high readings were recorded at Chai and other loacations. These can be viewed by moving the mouse over these points of interest.

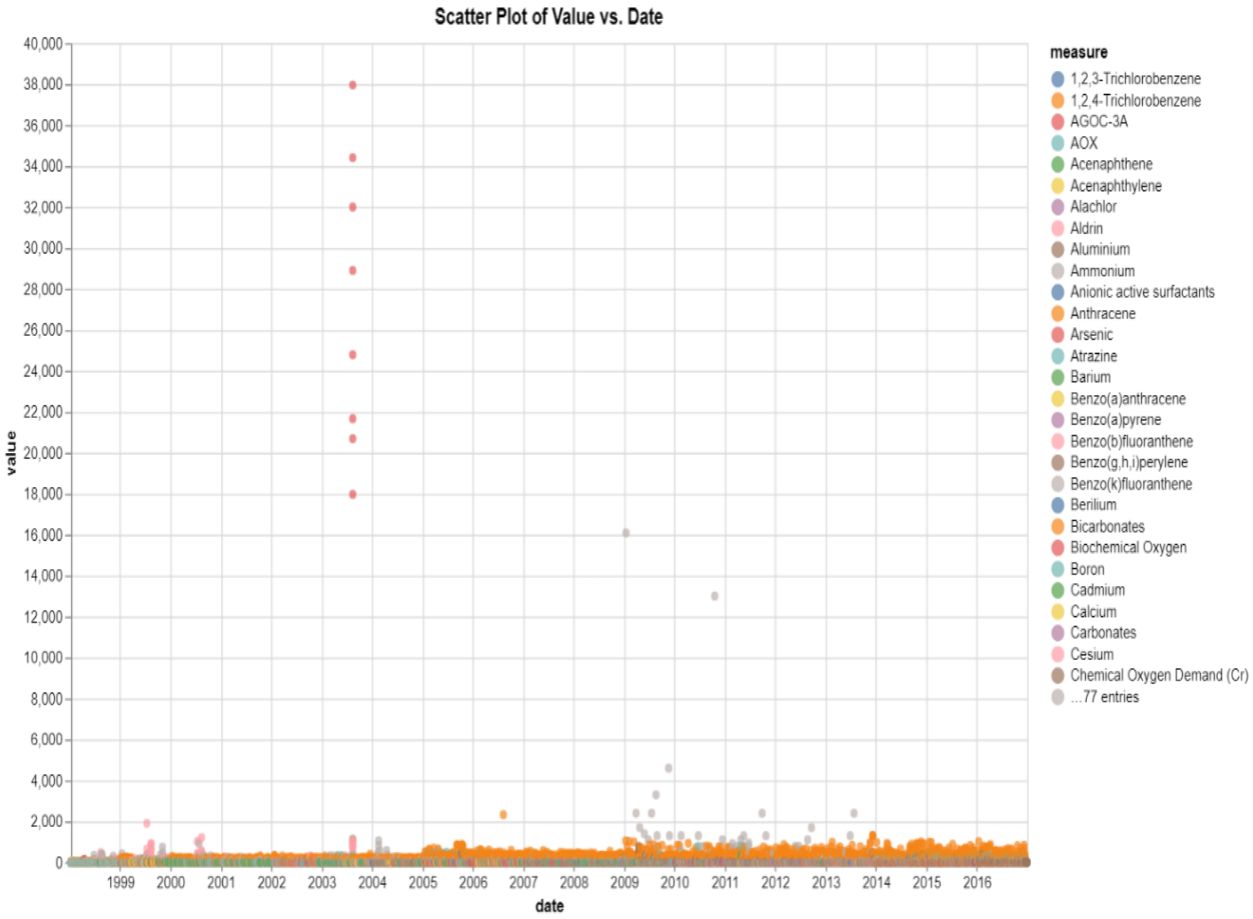


Figure 2 : Scatter plot showing unrealistic values of some measures

These high-risk values have been further shown with the boxplot below.

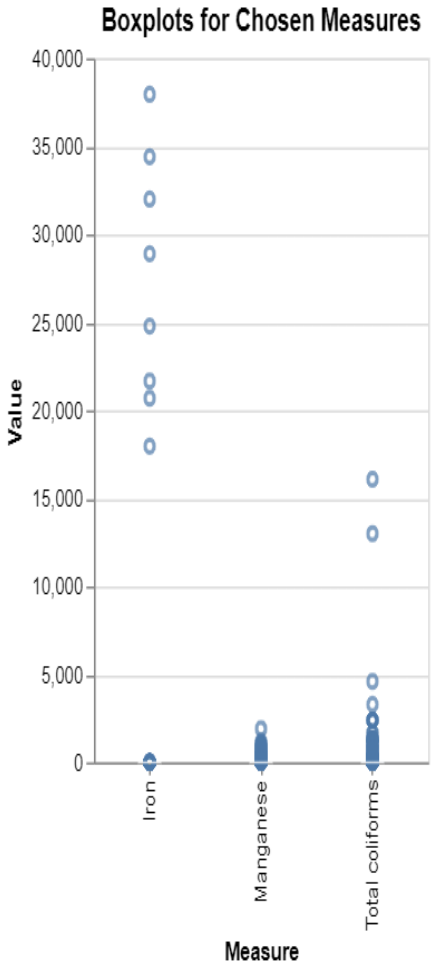


Figure 3: Box plot showing unrealistic values of iron, manganese and total coliforms.

2. Trends and Anomalies

- For Scalability, line charts of the measure values have been split to four years; 1998 – 2002, 2004 – 2008, and 2011 – 2016 as seen below.

- The bar chart in figure has been used as a benchmark in explaining the trends of different chemicals as it gives the average value of each reading.

- Line charts has been used to show trends and anomalies as it properly tracks changes in the readings over time.

1998 – 2002

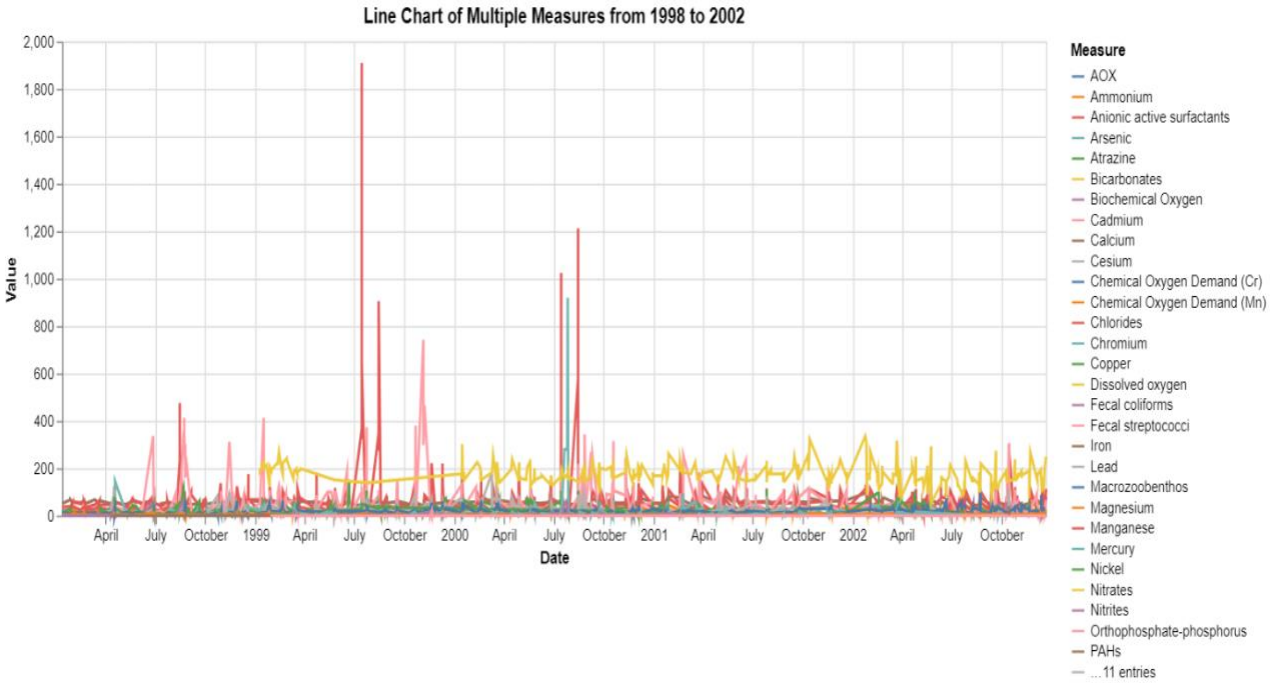


Figure 4: Line Chart of Multiple Measures from 1998 to 2002

Findings

Trends: From 1998 – 2002 as shown, most chemicals record the same range and behave in similar fluctuating manner except some points where visible spikes were observed which is further explained under the Anomalies sub-section.

Anomalies: During this time, we can see the spikes in manganese, zinc and coliform recording unusual values and showing the presence of these substances in water to be high.

2004 - 2008

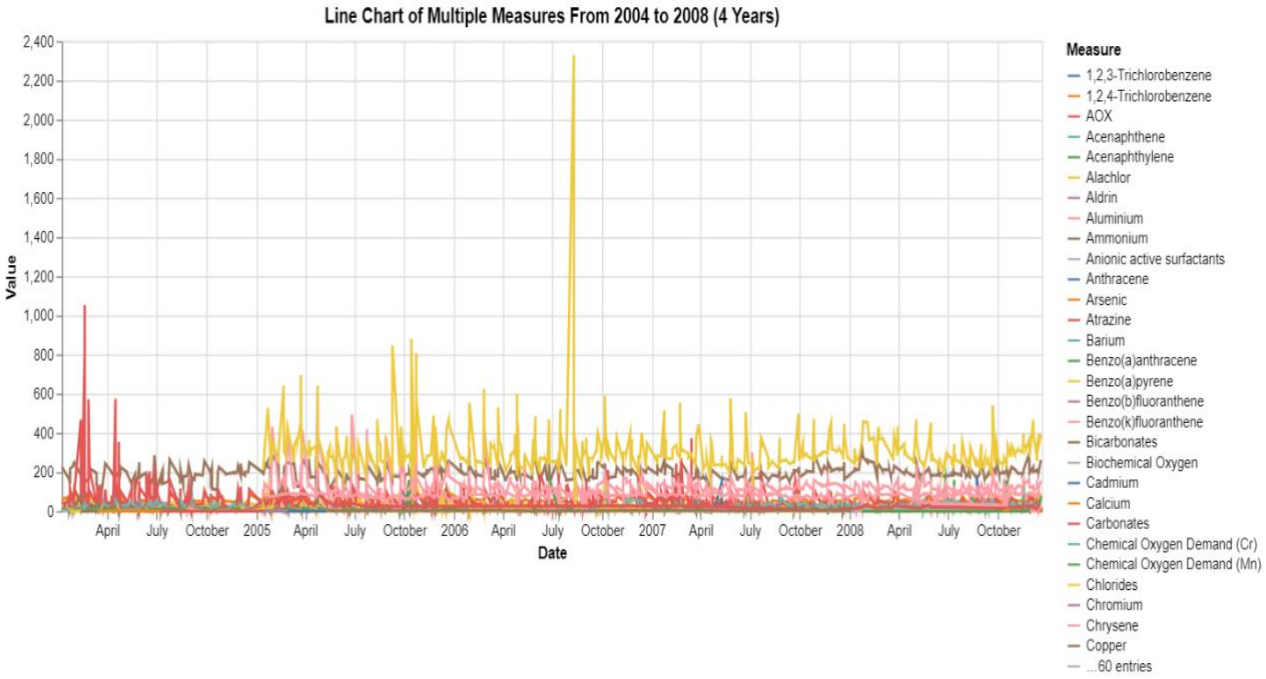


Figure 5: Line Chart of Multiple Measures from 2004 to 2008

Findings

Trends: From 2004 – 2008 as shown, most chemicals record higher readings when compared and behave in similar fluctuating manner except some points where visible spikes were observed which is further explained under the Anomalies sub-section.

Anomalies: This shows a spike in the start of the graph for zinc and then another spike for an abnormal value of dissolved salts at 2330.

2011 - 2016

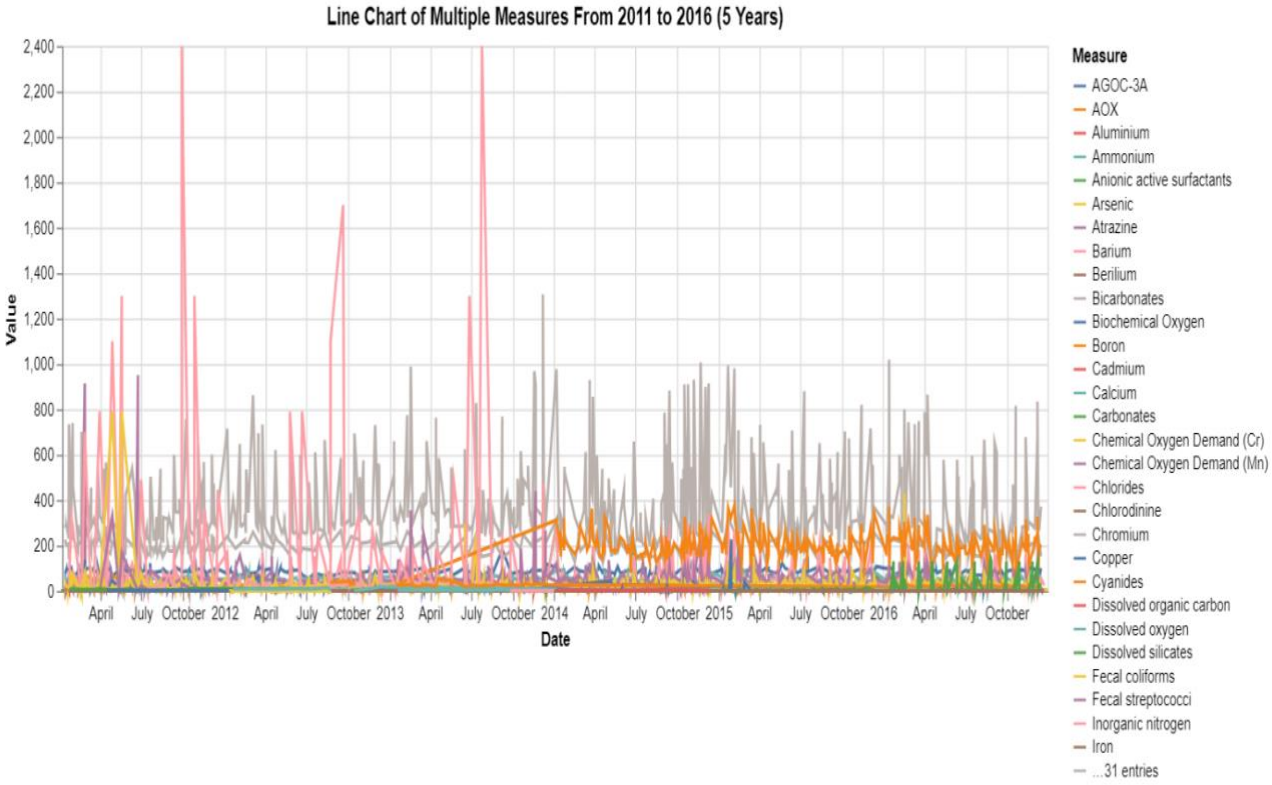


Figure 6: Line Chart of Multiple Measures from 2011 to 2016

Trends: From 2011- 2016 as shown, most chemicals record more spikes when compared to figure 5 and 6and behave in similar fluctuating manner except some points where visible spikes were observed which is further explained under the Anomalies sub-section. Also for total dissolved salts, from January 2013 – January 2014, there was a constant increase in the trend as against the usual fluctuations.,

Anomalies: this shows a high and irregular values of coliforms till 2013 and on one occasion a spike in dissolved salts shooting all the way up to 1300.