Over all in the report, 'Locations' will be inside an inverted commas and Measures will be underlined.

First, we will begin by answering the second analysis question which was about the quality of the data, there are 3 sub questions present in that.

Q2 (i) – Missing Data – To solve this, a pivot table was created so that for the missing values will have a 'Nan' instead of that observation not even being there. The nans will help us answering this question, further it has again melted into a long form from a wide form to retain the NaN's and getting a DataFrame that can used for a heatmap. NaN's are the empty white spaces in the heatmap below, indicating Missing Data.

The Finding – This graph allows us to know which measures have missing values for which years that shows us the quality of the data as well also helps us looking at the values of each measure that can help us answer future questions.

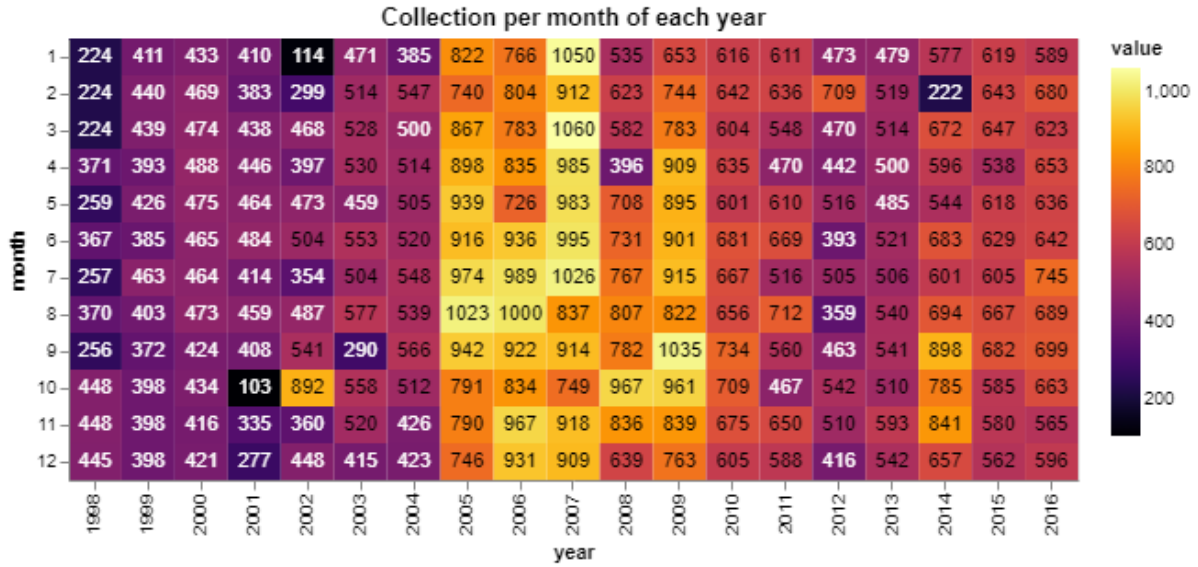
The What – The data here we have is in a tabular form (Pivot Table). The attribute here in the X axis 'Year' is a 'Sequential' datatype from Ordering Direction and on Y axis we have 'measures' that is 'Nominal' from Categories. The value of each measure is a Quantitative DataType.

The Why – In this graph we are looking for missing values of each measure for each year, even if a single value is present in 1 entire year, we will be able to see it or else it will be empty. In Action 'Analyze' I am 'Consuming (Discovering)' by finding new knowledge that was unknow before. In 'Produce' I am 'Recording' the graph which provides us the information we are seeking. In 'Search' I am 'Exploring' as I do not know which measures has missing values and how many of it is missing.

In ‘Query’ I am ‘Summarizing’ the full set of possible targets. In ‘Target’ I am looking at the ‘Features’ of all the data, for attribute I am looking at the lower end of the extremes how many times a component has been measured.

The How – I have 1 marker that is rectangles that will be filled with colour (1 Channel) from a gradient that represents the value of the given measure at a specific year. For this task a heatmap is the best choice as it allows us to have all the measures in 1 axis (Y axis in this graph) and all the years in the other axis, in addition to that we can also plot the value of each measure with each year, all of this can be done in a single graph and that’s why heatmap was chosen for this task. Using any other graph type will make us end up having 106 different visualisations or a visualisation with an issue of overplotting and humans can’t compile that much information all at once.

Q2 (ii) – Change in frequency of collection data. – After looking at the graph above, we have decided to remove all the measures that have in total of less than 180 recordings over the entire 19 years as they won’t have any impact on the collection frequency and on our analysis.



The Finding – This graph allows us to know the number of total recordings made in each month of each year that will help us answering the change in collection frequency over the period. Till 2007 from the beginning an increasing trend can be seen, post that an opposite trend is noticed. The Peak year for the number of collections was 2007

The What – Here we have a used a groupby function on the original data and obtained the results in a tabular form. The attribute here in the X axis ‘Year’ is a ‘Sequential’ datatype from Ordering Direction and the other attribute here in the Y axis is ‘Month’ that is ‘Cyclic’ datatype from Ordering Direction. The count is a Quantitative datatype.

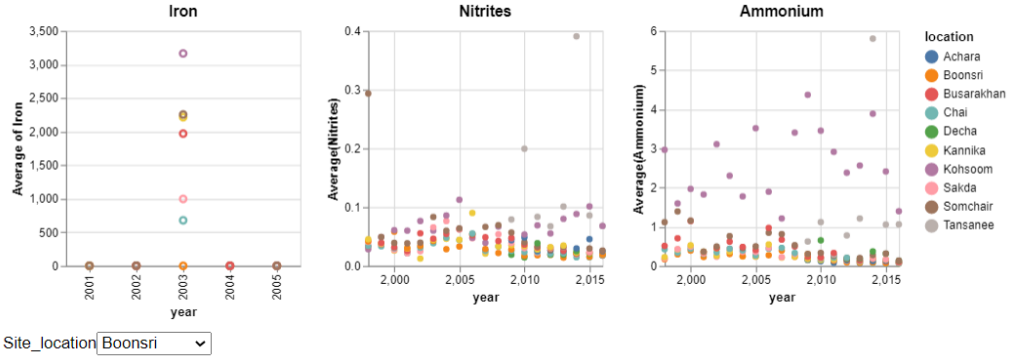
The Why – In this graph we are looking at the change in the frequency of collection of recordings for various measurements over the years. In Action ‘Analyze’ I am ‘Consuming (Discovering)’ by finding new knowledge that was unknow before. In ‘Produce’ I am ‘Annotating’ the number of recordings taken by the sensors. In ‘Search’ I am ‘Browsing’ as I do not know if there is any trend or change in the collection frequency or not. In ‘Query’ I am ‘Summarizing’ the full set of possible targets. In ‘Target’ I am looking at the ‘Features’ of all the data, for attribute i am looking at the similarities of each year’s collection amount via a heatmap.

The How – I have 1 marker that is a rectangle, and it will be filled by 1 Channel for encoding that is colour which indicates the number of times collection was made in each month of each year. In 'Reduce' I am aggregating the dataset using a groupby function to get the count of number of observations for each year and month. Again, here using a heatmap is the best choice as there is no other chart that would allow us to visualise such a vast amount of information in a single graph.

```
    color= alt.condition(datum.value> 500, alt.value('black'), alt.value('white')),
tooltip=['year','month','value'])

temp+text
```

For the heatmap above, 2 advanced features have been used. 1st layering, where 'temp' is the heatmap with only colors and 'text' is a marker which annotates our values. Layering them will give us a heatmap with colors+values. 2nd condition color, if the value is greater than 500 the text color will be black otherwise white (color black helps us seeing the text better in brighter shade of the gradient).

Q2 (iii) – Unrealistic Values – With the help of heatmap for missing data, we saw a measure called Iron, along with that I took at look at the 4 most frequently recorded measures. Nitrites and Ammonium are the ones having unrealistic value for the location 'Tansanee'.



The Finding – This graph allows us to see the unrealistic values recorded by the sensors. these values have been labelled as unrealistic because these values appeared out of nowhere in a recording and fell back down the following recording.

The What - The data here we have is the original dataset in a tabular form. The attribute 'Year' is a 'Sequential' datatype from Ordering Direction, the attribute 'Value' is a 'Quantitative' datatype.

The Why – The aim of this visualization is to look spot unrealistic values, these very well could be due to a faulty sensor. In Action 'Analyze' I am 'Consuming (Discovering)' by finding new knowledge that was unknow before. In 'Produce' I am 'Recording' the graphs. In 'Search' I am 'Locating' as I know that I want to find the measures with a sudden increase in value without any proper growth, but those measures are unknown (Location Unknown and Target known). In 'Query' I am 'Identifying' the outliers and using some prior data to compare if it's a real value or not. In 'Target' I am spotting 'Outliers' for specific measures and for attributes I am looking at the 'Extremes' of specific measures.

The How – Here we have 3 different charts but all of them share the same 1 marker that is point and 1 channel that is a color which represents various locations. Here usage of this graph is done to show that there is no gradual increase towards the unrealistic value, line graph will show a line going up towards the value which will not be clear if there were any other values present in between them, point shows the direct jump from one value to the another.

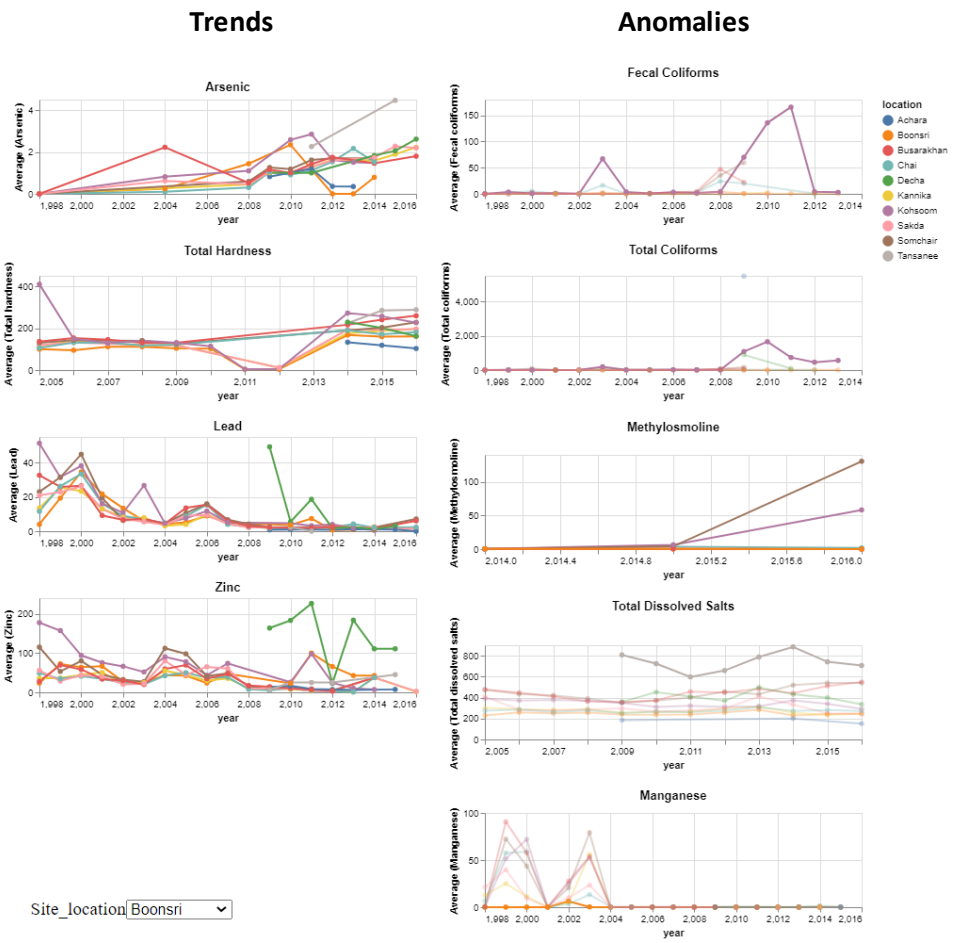
```
# Creating a drop down menu for selecting a site that will be used in future plots.
input_dropdown = alt.binding_select(options=df['location'].unique())

selection = alt.selection_single(fields=['location'], bind=input_dropdown, name='Site')

color = alt.condition(selection,alt.Color('location:N'),alt.value('lightgray'))

opacity = alt.condition(selection, alt.value(1.0), alt.value(0.2))
```

Here a drop-down menu has been created and used to filter through each location and apart from the selected location, other datapoints will be turned into light grey color with 0.2 opacity. Here the selection of drop-down menu will only affect 2nd and 3rd chart as they have various data points lying around the spike of value and only the location 'Tansanee' has seen this spike in Nitrites and Ammonium whereas for Iron almost all the location suffered.

Now we move on to the First Analysis Question that we were given.

Q1 (i) Find Trends, (ii) Find Anomalies.

We will answer these questions together. To answer this question first we create a pivot table with each measure as Index and each year as column and the mean value of that measure for that year. After that we will use the diff() function on columns and we will get is the difference of each measure's value from its previous year's value, as long as the values are close to each other and not only going in 1 direction, we can consider them being normal. We note the ones that are special. After noting them down, we explore them further and separate them into correct sections.



The Finding – Here on the left we have all measures that have an increasing or decreasing trend and the right we have all the measures that consist of an anomaly.

In 'Trends' Arsenic and Total Hardness have an increasing trend, although Lead and Zinc has noticed an opposite trend.

For 'Anomalies' location 'Kohsoom' is different than other sites in terms of Fecal and Total Coliforms, in addition to that 'Somchair' is also different for Methylosmoline. 'Tansanee' is different than other sites when it comes for Total Dissolved Salts. 'Boonsri' does not follow the Zig-Zag patterns as other sites for Manganese in the beginning few years.

The What – Here we have our data in a tabular form (Pivot Table) where the attribute 'Year' is a 'Sequential' datatype from Ordering Direction, the attribute 'Value' is a 'Quantitative' datatype.

The Why – This group of visualizations/Mini-Dashboard aim is to show us the measure's having trends and anomalies. In Action 'Analyze' I am 'Consuming (Discovering)' by finding new knowledge that was unknow before. In 'Produce' I am 'Recording' the graphs. In 'Search' I am 'Exploring' as I want to find the measures that have a trend/anomaly. In 'Query' I am 'Comparing' the values of each measure with each year. In 'Targets' for all data, 'Trends and Outliers' is what I am after. In 'Targets' for attributes I am looking through 'many' measures and finding 'Similarites' throughout the years. If the similarity is present then it can be considered as a trend, if not then it could be an anomaly.

The How – I have 2 markers combined here and those are a line and point. Line helps us see how the values are flowing overtime and points help us compare the values for a specific year of various locations by just looking vertically. Along with that I also have 1 channel that is color which represents each location. All the Graphs share a legend among them. A line graph is the best in terms of looking at data over the years or any given time. Other graphs make it difficult to follow trends of various locations in a single graph. For trends there is a drop-down menu which allows us to focus only on 1 location at a time if we want to (Decha for Lead and Zinc could be an Anomaly as well). For Anomalies, the opacity has already been reduced to 30% for all the other locations than the one having an Anomaly. Also, there is a transform filter at the end of each graph to filter out the data for only 1 specific measure at a time.

```
color='location:N',
opacity=alt.condition(datum.location=='Tansanee', alt.value(1), alt.value(0.3)),
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Total dissolved salts').properties(height=100, width=400, title='Total Dissolved Salts')
```

This is an example of Anomaly measure 'Total Dissolved Salts' and it has been filtered and the opacity has been reduced for all the locations but Tansanee.

A tooltip has been added to all the graphs which shows information such as the date/year of recording and the value itself including the location also. Use of transform filters to select individual measures has also been used in all the graphs for trends, anomalies, unrealistic values.

Concatination has been used for the graphs of Unrealistic values, Trend and anomalies.

Other Information - There were few other measures such as Atrazine, gamma-Hexachlorocyclohexane, Chlorodinine which showed a high positive/negative correlation with time but due to inconsistency in collection frequency and there being huge gaps between the years of collection. They cannot be considered for a trend.