Vega-Altair – Python Library – Environment Analysis.

Nishita Chaudhary

Dataset information: -

The Boonsong_Lekagul.csv is a tabular dataset which have 136824 rows and 5 columns that include id (int), value(float), location, sample date and measure (object respectively). There are 10 locations. There are 106 measures.

Attributes	Attribute Type/Encoding Dataset	Explanation
Value (Measures)	Quantitative (Q)	Values which can be calculated to perform mathematical operations
Location, measure (Dimensions)	Nominal (N)	Categories
Sample date	Temporal (T)	Gives the data time

Data Analysis: -

The dataset has no null values and there are no duplicates. However, after carefully observation in data quality chart some areas the data is not observed and recorded as 0. Th e dataset has year from 1998 to 2016.Using python Libraries pandas to import dataset. The chart type used are Marks : mark_line() , mark_bar(), mark_rect(), mark_point.

Trend: Chart-1

Marks: The marks in this chart are lines, each representing the trend of the average value over time for a specific location.

Channels:

X-axis (x): Temporal channel, representing the year.
Y-axis (y): Quantitative channel, representing the average value.
Color (color): Nominal channel, representing different locations with distinct colours.

Interactive: Added the selector (radio button) and transform_fiter that filters the data based on location. The parameter here used is selection_point for selecting each point of locations and widget used binding_radio for selecting the location using radio buttons.

Trend 1 –

The line graph shows clear trends in the average values for each place. Kohsoom saw a significant uptick in 2003 and a precipitous drop in 2004. That year, Somchair, Busara khan, and Kannika all showed comparable rises at the same time. Tansanee, on the other hand, started the highest average values in 2009 and continued to do so until 2016. T hese notable variations over time are clearly highlighted by the line marks on the chart, which stand for the average values.

```
trend_chart = alt.Chart(df).mark_line().encode(
    x='year(sample date):T',
    y='average(value):Q',
    color=alt.Color('location').scale(domain=options)).properties(
   width=600,
    height=300,
    title='Avearge of Value for Locations Across Year'
```

Fig 1.1 Code for Trend chart

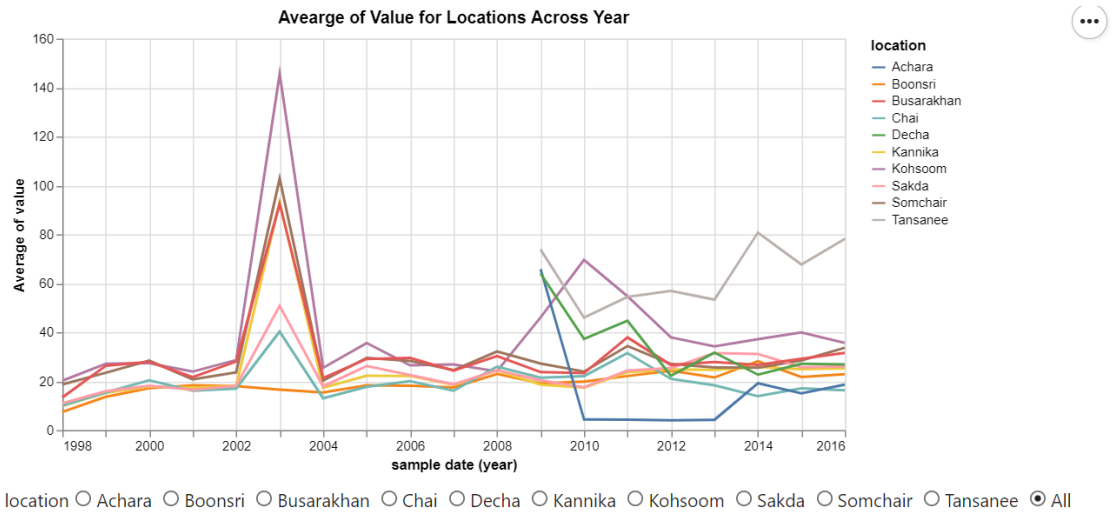

Fig 1.2 Average of value for all location across year.

Chart- 2

Marks: The marks in this chart are lines,that represents the trend of the sum of values over time for each of the top 4 measures. The bar chart uses marks of type 'bar' to repres ent the total sum of values for each measure.

Channels:

X-axis (x): Temporal channel in the line chart, representing the month and year. Nominal channel in the bar chart, representing the measure.
Y-axis (y): Quantitative channel in both charts, representing the sum of values.
Color (color): Nominal channel in both charts, distinguishing measures by color. Conditional colouring based on the selection.

Interactive: Added the selector_point that creates a point selection of measures and transform_fiter that filters the data based on measure.

Trend 2-

The total of the values for the top four measures shows the temporal patterns in the line chart. The total value for each of the top 4 metrics is shown in the bar chart, which off ers a summary view. The analysis considered the top four measures with Total Dissolved Salts showing the highest cumulative sum of values. Since the observation began in 2005, Total Dissolved Salts has continuously maintained the highest values among the chosen measures. Particularly, aluminium displayed a unique pattern: it saw a sharp in crease in early January 2009, then a gradual decrease that ended in mid-2015. Total dissolved salts possibly the key factor affecting the river. In order to filter both charts at o nce, users can interactively choose a measure, which allows for a targeted analysis of particular measures throughout time.

```
top_4 = alt.Chart(top_4_df).mark_line().encode(
    x='yearmonth(sample date):T',
    y='sum(value):Q',
    color='measure:N').properties(
    width=600, height=300,title='Sum of Value for measures Across MonthYear'
).transform_filter(
    selection)
#Creating a bar chart of top 4 measure
bottom = alt.Chart(top_4_df).mark_bar().encode(
     x='measure:N',
    y='sum(value):Q',
    color=alt.condition(selection, 'measure', alt.value('lightgray'))
).properties(
    width=600, height=100
).add_params(
    selection
)
```

Fig 1.3Average of value for all location across year.

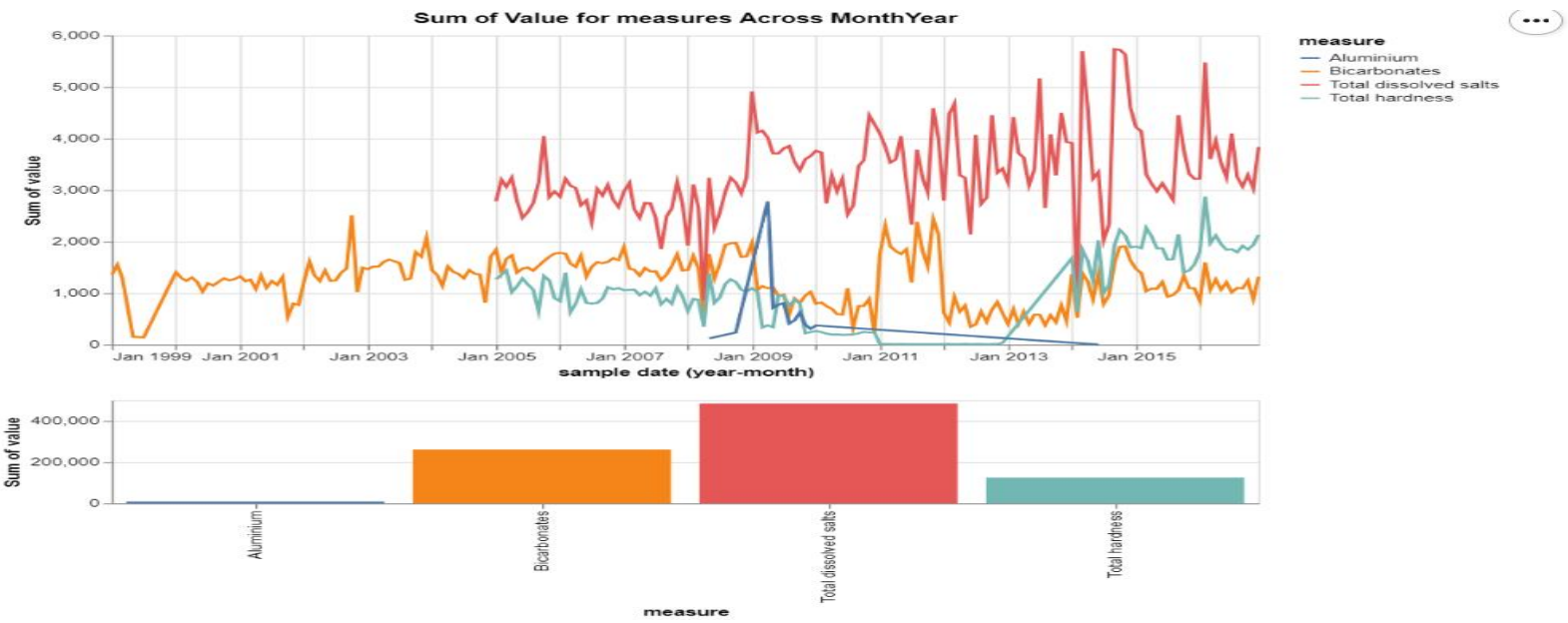

Fig 1.4 Sum of value for top 4 measure across monthyear

Anomalies:

Marks: The marks in this chart are lines, that represents the trend of the sum of values over time for each of the top 4 measures. The bar chart uses marks of type 'bar' to repre sent the total sum of values for each measure, with conditional colouring based on the selection.

Channels:

X-axis (x): Represents the year in the line chart and count in the bar chart.
Y-axis (y): Represents the average value in the line chart and the location in the bar chart.
Color (color): Represents different measures in both charts.

Interactive:

Brushing: Using the brush, users can interactively choose different parts of the line chart.
Radio Button Selection: The radio button allows users to choose particular measures.

Anomalies Chart:

Selecting only measures that include the total quantity, such as "Total Coliforms" and "Total Dissolved Phosphorus, “etc. To get an accurate measure result of affecting the ri ver. Total coliforms suddenly changed from 2008 to 2012; it increased from 2008 to 2010 and then significantly decreased for the next two years. Talking about total hardnes s, it started from 2005 to 2010; it slightly decreased over these years and had a sudden drop for one year, then was steady from the next year and unexpectedly increased for t wo years, then slowly increased till the end of the year. Total dissolved salts is the key chemical effecting the river. The location with the most affected can be seen in Boonsr i. A selection interval on the graph allows the selection of a particular area, and a related bar chart that shows the count of locations provides insight into the distribution of da ta within that chosen range.

```
#Creating a line chart with year and average value having color as measure
anomalie_chart = alt.Chart(filter_df).mark_line().encode(
    x='year(sample date)', y='average(value)',color='measure:N',
).properties( width=600, height=300, title='Average value for selected measure across year'
).add_params(brush)
#Displaying the line chart
anomalie_chart
```

Fig 2.1 Code for Anomalies line chart.

```
#Creating a bar chart which will dispaly the count of records with radio button when selection on the line graph
anomalie_bar = alt.Chart(filter_df).mark_bar().encode(alt.X('count()'), alt.Y('location'), color = "measure"
    ).transform_filter(brush).properties(width=600, height=300,title='Count of records with location over Selected Measures'
).add_params(selector).transform_filter(selector)
```

Fig 2.2 Code for Anomalies bar chart

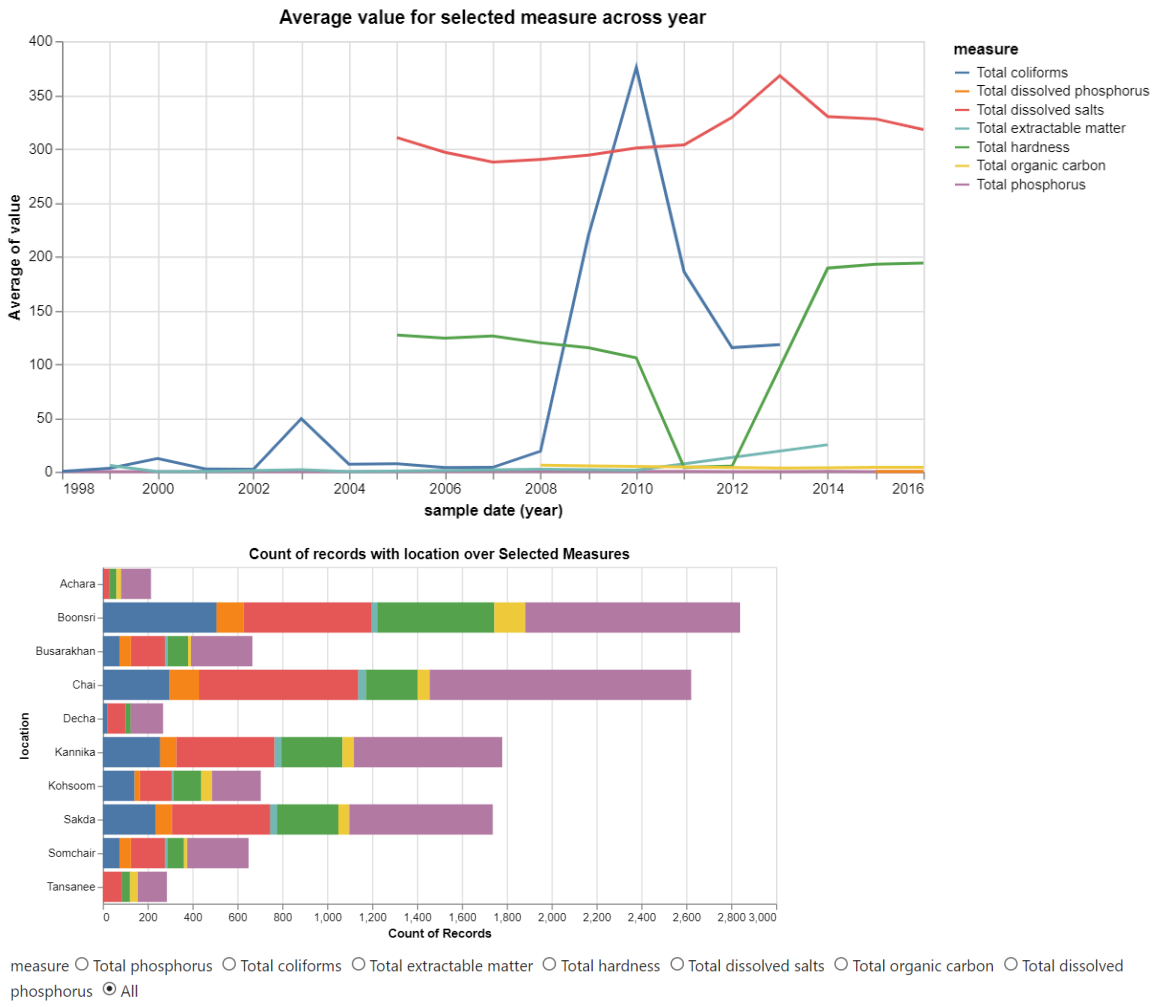

Fig 2.3. Average of value for total selected measure across year.

2. Describe any data quality and uncertain issues

Marks: Circle marks are used in both scatter plots.
Channels:

X-axis (x): Represents the year as a temporal field.
Y-axis (y): Represents the numerical values.

Removal of outliers:

The data point which deviates significantly from the dataset are considered as outliers. Most deviated change can be seen in 2003 the data point went between 15,000 to 40,00 0 that have the largest point. These outliers can be removed by IQR in pandas. Potential outliers are defined as data points that fall outside of Q3+1.5IQR or outside of Q1-1.5 IQR. After removing the outliers 18,844 rows are removed. The original data is shown in the first graphic (data_original), with red highlights denoting the locations of outlier s. After removing outliers, the data is displayed in the second chart (data_no_outliers), which offers a more accurate representation of the data distribution.

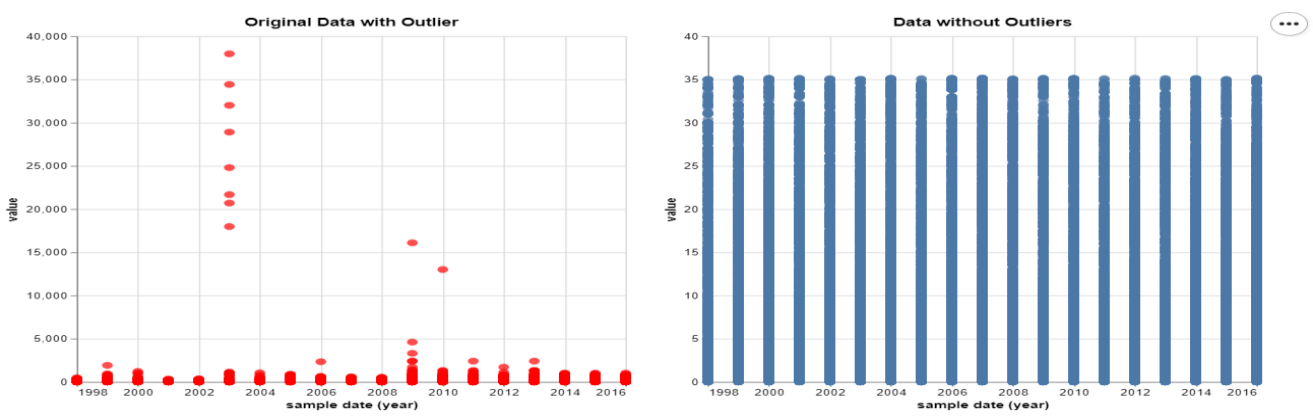

Fig 3. With and without outliers' data chart

i. Missing data –

Chart 1

Marks: marks is used mark_bar()
Channels:

X-axis (x): Represents the location as a Nominal field.
Y-axis (y): Represents the count of zero values as Quantitative field.
Color (color): Nominal channel in bar charts, distinguishing locations.

Interactive: Opacity

Missing chart

Boonsri has more zero values than any other region (more than 2,000 records), which may have a major effect on the distribution of data. Tansanee, on the other hand, shows t he fewest zero values, suggesting a more varied collection of non-zero data points. Given that these areas are largely composed of zeros, it is important that preprocessing be d one with carefully to lessen the impact of these values on any other observations.

```
chart_0_values = alt.Chart(zeros_count).mark_bar(opacity=op_var).encode(
    x='location:N',
    y='countt:Q',
    color=alt.Color('location:N')).properties(title='Count of zeros in location',
    width=600,
    height=200).add_params(op_var)
chart_0_values
```

Fig 4.1 missing data code

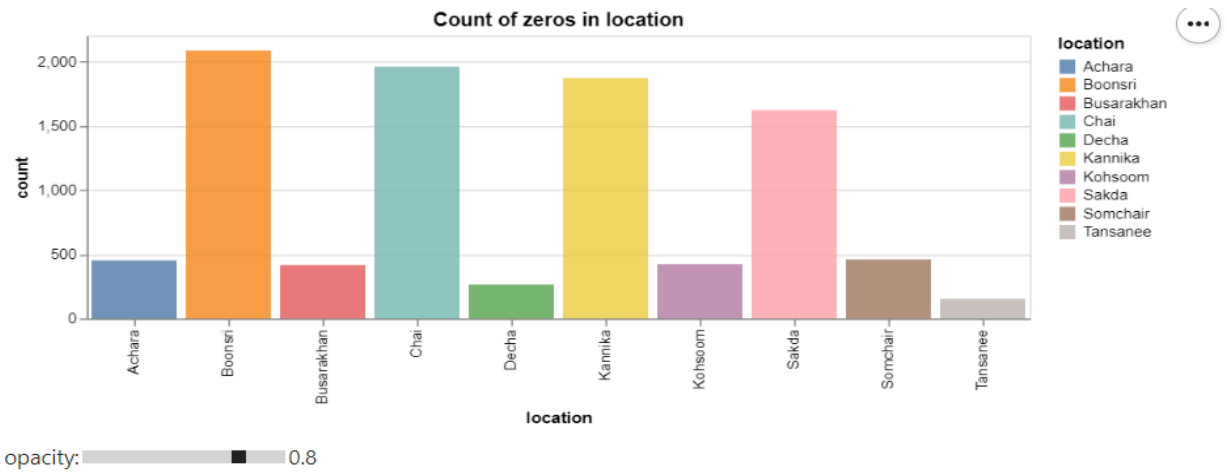

Fig 4.2 Count of zeros in each location.

Chart-2

Marks: Rectangular marks are used in heat map (mark_rect).
Channels:

X-axis (x): Represents the sample date as a Temporal field.
Y-axis (y): Represents location as Nominal field.
Color (color): Conditional colour encoding

Interactive: Selection point for selecting the particular box in heatmap and show the value.

Missing chart

The heat map represents, From 1998 to 2009, all location indicates values that are continuously below 0.1, according to the heat map. Achara, Decha, and Tansanee don't see m to have much data during 1998-2009, which could indicate that the dataset for these places during these years has either missing or insignificant data. This finding highligh ts insufficient information or missing data in Achara, Decha, and Tansanee over the given time. In other areas as well, the data appears to be near zero. The light-yellow bloc ks in heat map indicates that the value is 0.0.If selection one block the other blocks turns grey giving the particular location in a year insight with a values using tooltip.

Fig 4.4. Heat map Count of zeros in each location.

ii. Change in collection frequency.

Marks: Mark bars are used for change in collection of frequency by overall, monthly, weekly and quarterly.
Channels:

X-axis (x): Represents the Year as Temporal, month, week, and quarter as a Nominal field.
Y-axis (y): Represents the count of dates as Quantitative field.

Frequency chart.

A multi-granular view of event frequencies across time is possible with the combined visual.Charts that display the distribution of events across longer time periods on a mon thly, weekly, and quarterly.The yearly figure shows a 2007 peak, showing a significant rise in the frequency of events that year. Consistent monthly patterns indicate counts a bove 10,000, with months 6 and 8 seeing the highest activity. Weekly variations show that counts are highest in week 50 and lowest in weeks 5 and 9.
The third quarter has the largest frequency of incidents that are recorded, according to the quarterly chart.

```
#Making frequency charts for the day, month, week, and quarter
daily_chart = alt.Chart(df).mark_bar().encode( x='sample date:T', y='count():Q').properties(width=550,height=150,title='Frequency of overall Events')
monthly_chart = alt.Chart(df).mark_bar().encode(x='month:N',y='count():Q').properties(width=200,height=150,title='monthly Frequency')
weekly_chart = alt.Chart(df).mark_bar().encode( x='week:N',y='count():Q').properties(width=500,height=150,title='weekly Frequency')
quarterly_chart = alt.Chart(df).mark_bar().encode(x='quarter:N',y='count():Q').properties(width=200,height=150,title='quarterly Frequency')
#Combining charts with vertical and horizontal concatentation
overall_chart = alt.vconcat(daily_chart, alt.hconcat(monthly_chart, weekly_chart, quarterly_chart))
```

Fig 5.1 code for count of frequency .

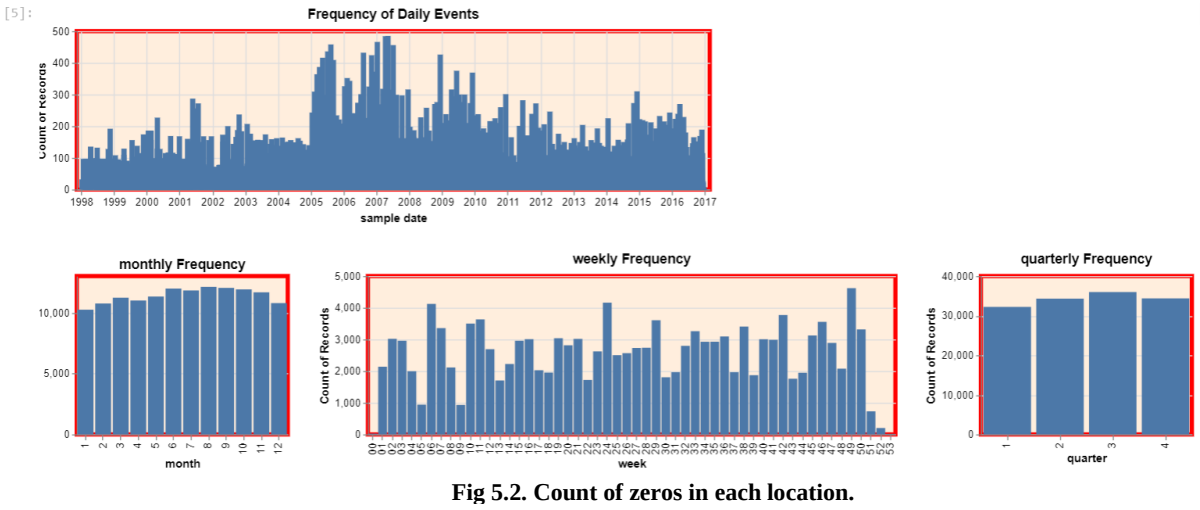

Fig 5.2. Count of zeros in each location.

iii. Unrealistic values.

Chart-1

Marks: Line marks are used to represent the trend of water quality measures over time.
Channels:

X-axis (x): Represents the sample date as a temporal field.
Y-axis (y): Represents the measured values.
Color (color): Differentiates measures.

Interactive: Dropdown Selection (brush) allows users to interactively select water quality measures from the dropdown, updating the displayed charts accordingly.
Unrealistic Chart-1

The graph shows the contamination of water properties in river that include dissolve oxygen, oxygen saturation and water temperature. The oxygen saturation graph in the first chart shows wild fluctuations, but around 2006 there is a least trend. Around 2007, there is also a sharp rise. A significant change is seen after the elimination of the outlier. After showing a wide range at first, the values now display a more limited range between 10 and 35. With this correction, the impact of outliers on the original analysis is highlighted. The chart uses dropdown to select the measures. By comparing water quality measurements with and without outliers, we can identify unrealistic numbers with the help of this visualisation.

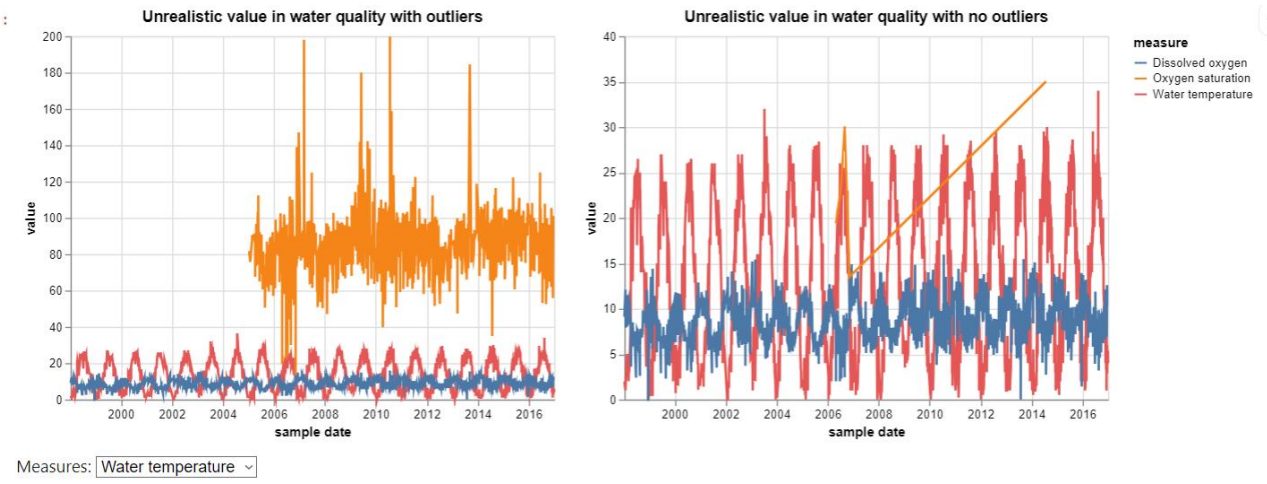

Fig 6. Sum of value for top 4 measure across monthyear

Chart-2

Marks:

Point marks are used in the water temperature chart to represent average values.
Point marks in the legend represent each location.
Rule mark is used for the red line indicating the overall average water temperature.

Channels:

X-axis (x): Represents time (Year-Month) in the water temperature chart.
Y-axis (y): Represents the average water temperature in both the chart and the rule.
Color (color): Encodes different locations conditionally in the chart and legend.
Tooltip: Provides additional information on the chart and legend.

Interactive: Selection (selection) enables interactive highlighting of locations when points are selected.

Unrealistic Chart-2

The graph shows the trends in average water temperature throughout time at different places.The overall average water temperature is determined using the red line rule as a guide. From 0 to 30 degrees is the typical range and pattern shown in the average water temperature chart. There is, however, one exception, which is in "Kohsoom," where the average water temperature reaches an exceptionally high 36.4 degrees. And Tansanee reached to 34. We can target particular locations of interest with interactive selection.

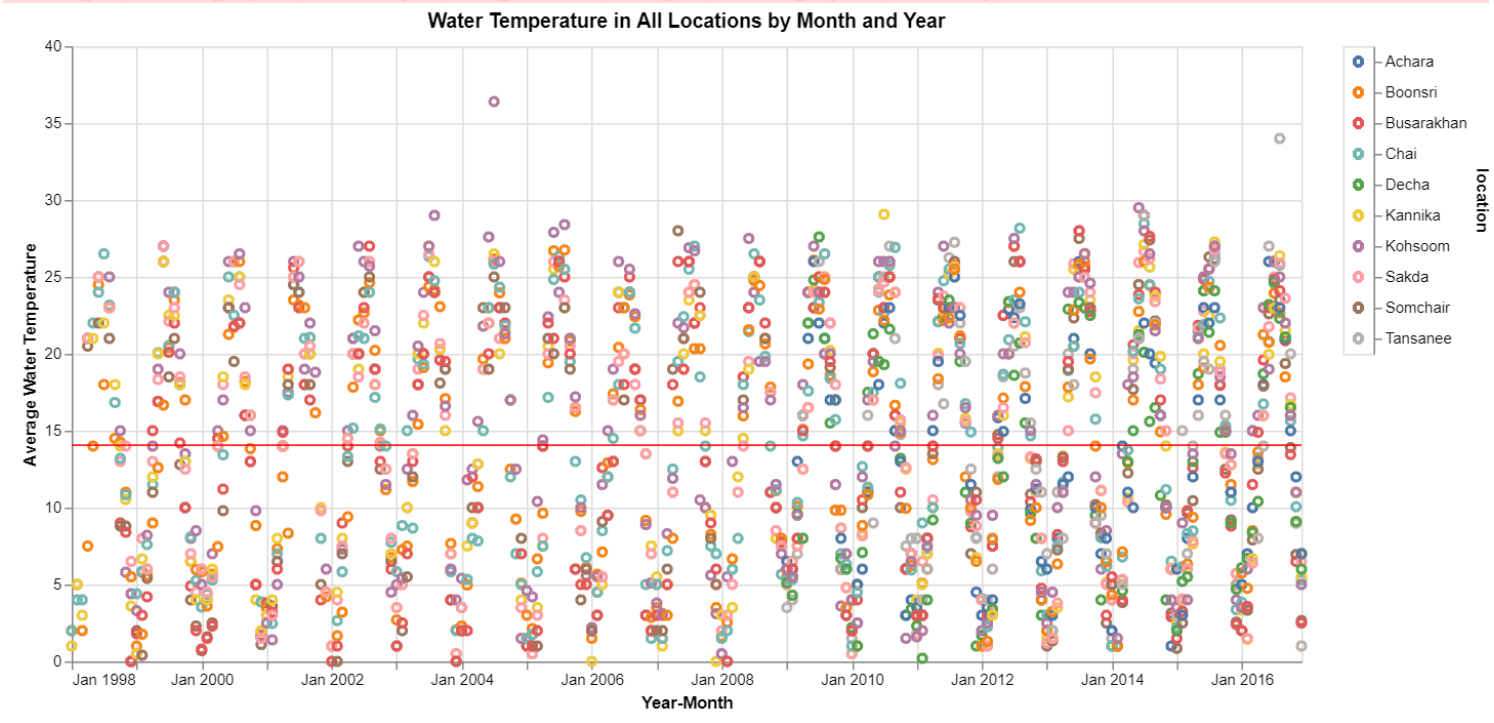

Fig 7. Average water temperature in all locations by month year