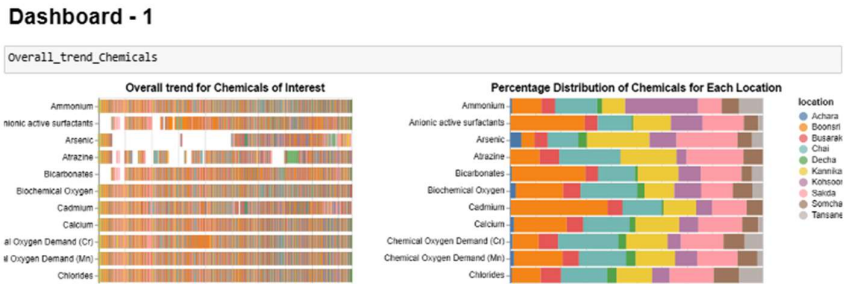
COURSEWORK -2

VISUAL DATA ANALYSIS – ALTAIR

TREND AND ANOMALIES WITH RESPECT TO CHEMICAL CONTAMINATION

This coursework is to provide contamination study based on 19 years of data which includes 106 different chemicals spread across 10 different locations. Chemicals which have very few readings and does not show interesting trends over time is of least importance hence such chemicals are excluded from this analysis. Based on this, 41 chemicals are considered for the overall trend.



Only part of visuals is included to save space on explanation.

Findings

- **Overall trend of chemicals** – Shows distribution of chemicals over the years. Some chemicals have values for all the years, and some are missing data for some years.
- **Percentage distribution of chemicals** - Achara, Decha and Tansanee have lower percentages compared to other location due to data collection delay which started only from Y 2009.
- Large percentage of chemical for a particular location may not indicate large contamination. Kohsoom showed higher percentage for Total coliforms. This may not indicate actual contamination. Higher values might be also due to outliers.

Effectiveness of Visual encoding

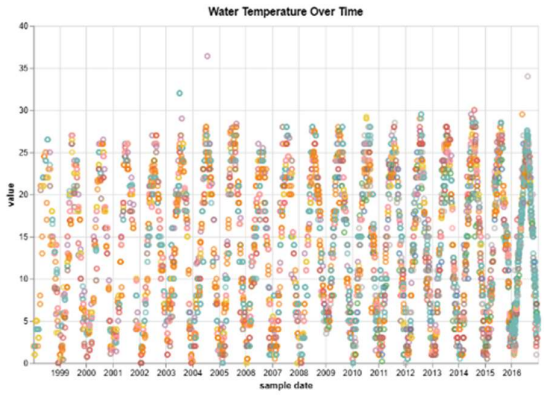
- **Mark1 - Tick chart** to show the overall trend of chemicals over years. They are highly granular and individual datapoints can be clearly seen in the chart.
- **Mark2 - Stacked bar chart** to show the Percentage distribution of chemicals for all waterways. They show clear separation between different waterways. It is easily interpretable.
- **Channels –** position (x-year, percentage; y-measure), color- to highlight values based on locations, tooltip to show information when hovered over the datapoint.
- **Datatypes**
  i. Temporal: sample date - used for encoding time-based information.
  ii. Nominal: location - used for encoding categorical values without any order.
  iii. Quantitative: value – used for encoding continuous values in visualization.

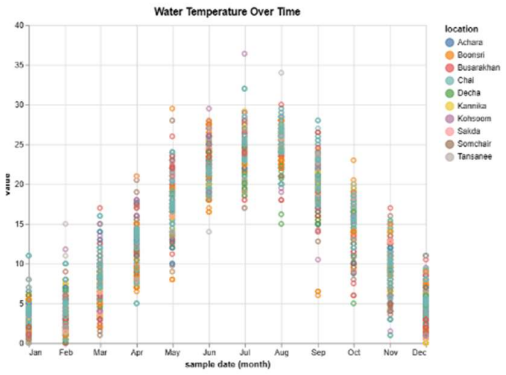
Explanation of the Altair code

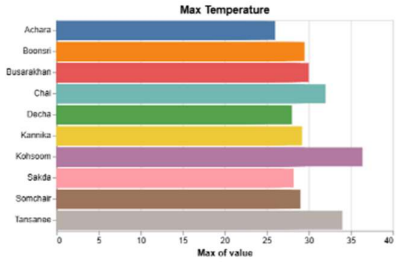
List of chemicals are created and filtered into a dataframe and then tick chart is created to show the overall trend. Percentage calculation is done for the bar chart using below code snippet. Transform function is used here as they do not reduce the number of rows when grouping instead it assigns the same sum for similar measure repeating the value over multiple lines. Horizontal concatenation is done for these two charts to form Dashboard-1

```
#calculating the percentage of chemicals for each waterways-(chemical in particular location/total chemical)*100
total_per_chemical = DF1.groupby(['location','measure'])['value'].sum().reset_index()
total_value_per_chemical = total_per_chemical.groupby('measure')['value'].transform('sum')
total_per_chemical['percentage'] = (total_per_chemical['value'] / total_value_per_chemical) * 100
```

```
#horizantal concatenation of tick and stacked bar chart
Overall_trend_Chemicals = alt.hconcat(Data_distribution,percentage_distribution)
```







Findings:

- Water temperature was at its peak for the month of June, July, and August for every year. This could be due to the release of industrial wastewater at these locations. As the contaminations were diluted and washed away, the temperature started reducing gradually by month end.
- Anamoly - Koshoom had high temperature of 36.4 degrees for the year 2004 which may indicate high level of contamination at this location for that year.
- Anamoly - Temperature for Tansanee was showing a very high value of 34°C for the year 2016 which may indicate high level of contamination in this location.
- Location Chai has frequent reading for the year 2016, this indicates that some regulatory bodies might be assigned to this location to monitor data collection.

Effectiveness of Visual encoding

- **Mark1 & Mark-2 – Interactive scatter chart** to show the distribution of water temperature over years and month. Using this chart, it was easy to observe the trend and datapoints clearly.
- **Mark3 - Bar chart** to show the max temperature from selected range of values. The values can be easily interpreted using the bar chat.
- **Channels –** position (x-sample date, max(value); y-value, location), color- to highlight values based on locations, tooltip to show information when hovered over the datapoint.
- **Datatypes - T**emporal: sample date, Nominal: location, Quantitative: value.

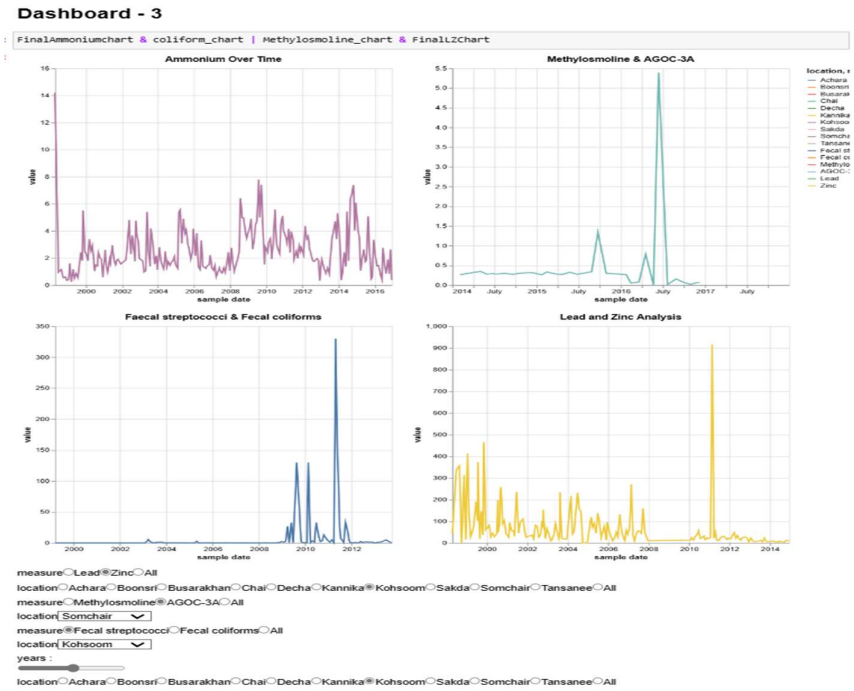
Explanation of the Altair code

Dataframe is created filtering the measure for water temperature. Brush variable is created to select the range of values for particular year. Max temperature is calculated from these selected values in another barchart making them interactive. Conditional statement in the color channel to highlight the selected values in color and rest of the values in light grey.

```
brush = alt.selection_interval(encodings=['x'])

Watertemperature_chart = alt.Chart(filtered_df).mark_point().encode(x='sample date:T', y='value:Q',
                                                    color=alt.condition(brush,'location',alt.value('lightgray')), tooltip=['sample
```

All the three charts are combined to form Dashboard-2.

```
Watertemperature_Trend = (Watertemperature_chart| Watertemperature_chart1) & (WT_bar1)
```



Findings: (Anomalies)

Ammonium – All the location showed decreasing trend except koshoom and Tanseen. Tanseen showed very high ammonium level in the year 2014. Koshoom showed high level of ammonia for most of the years. It also showed a sudden spike in the year 1998 which could be a measurement error.

Methylosmoline & AGOC-3A – They have inverse relationship. AGOC-3A is an environment friendly replacement to methylosmoline. With increase in AGOC-3A, methylosmoline value either decreased or it maintained a constant value.

1. At koshoom, the historic reading for methylosmoline did not show any variation but in the year 2015 it increased to a very high value with a spike and started decreasing by the end of 2016. This indicates that the company stopped dumping this chemical by the end of 2016.
2. At Somchai, in the year 2015 the value increased slightly and remained constant for certain period and then increased very high and remained constant for period. This might look like a very high contamination but it is actually an systematic measurement error.

Fecal streptococci & Fecal coliforms – These are microbial contamination in water. They are bacteria preset in the intestine of birds and animals. Presence of this chemical at any location shows that the birds have consumed the contamination from these waterways. This is turn reduced the bird's population at certain locations.

1. We have values only until 2013, we do not have the values for Achara, Decha and Tansanee.
2. Fecal streptococci & Fecal coliforms have positive correlation. When there is an increase in fecal streptococci there was an increase in faecal coliforms.
3. Koshoom had a very high value from 2009 to 2011which shows high contamination.

Lead & Zinc – Both showed decreasing trends for all locations in recent years. For koshoom and boonsri, the values are missing for the years 2008 to 2010. It might be deleted intentionally to maintain the value below certain threshold to have the decreasing trend as other locations.

Effectiveness of Visual encoding

- **Mark –** 4 different Interactive line charts to compare Ammonium, Methylosmoline & AGOC-3A, Fecal streptococci & Fecal coliforms, Lead & Zinc. I have used the line chart for better readability and highlighting the anomalies.

- **Channels** – position (x-sample date, y-value), color- to highlight values based on locations, tooltip to show information when hovered over the datapoint.
- **Datatypes -** Temporal: sample date, Nominal: location, Quantitative: value.

Explanation of the Altair code

Ammonium chart has radiobuttons to select the location and slider to select the year. Based on these values the chart is filtered for location and year.

```
options=['Achara','Boonsri','Busarakhan','Chai','Decha','Kannika','Kohsoom','Sakda','Somchair','Tansanee']
labels =[option +' '  for option in options]
bind_radio = alt.binding_radio(options=options+ [None],labels=labels+ ['All'],name='location')
select_radio = alt.selection_point(fields=['location'],bind=bind_radio)


year_slider = alt.binding_range(min=1998,max=2017,step=1,name='years : ')
slider_selection = alt.selection_point(bind=year_slider, fields=['Year'])
```

Methylosmoline & AGOC-3A Analysis chart has radio buttons for selected chemicals and dropdown box for locations. Values are filtered through these interactions and the line chart is shown.

```
#dropdown creation
dropdown1 = alt.binding_select(
    options=['Achara','Boonsri','Busarakhan','Chai','Decha','Kannika','Kohsoom','Sakda','Somchair','Tansanee'],
    name='location'
)
dropdown_select1 = alt.selection_point(fields=['location'], bind=dropdown1)

#radiobutton creation
options=['Methylosmoline','AGOC-3A']
labels =[option +' '  for option in options]
bind_radio2 = alt.binding_radio(options=options+ [None],labels=labels+ ['All'],name='measure')
select_radio2 = alt.selection_point(fields=['measure'],bind=bind_radio2)
```

Fecal Streptococci & Fecal coliforms chart has radio buttons for selected chemicals and dropdown box for locations. Values are filtered through these interactions and the line chart is shown.

```
#dropdown creation
dropdown = alt.binding_select(
    options=['Achara','Boonsri','Busarakhan','Chai','Decha','Kannika','Kohsoom','Sakda','Somchair','Tansanee'],
    name='location')
dropdown_select = alt.selection_point(fields=['location'], bind=dropdown)

#radiobutton creation
options=['Fecal streptococci ','Fecal coliforms']
labels =[option +' '  for option in options]
bind_radio1 = alt.binding_radio(options=options+ [None],labels=labels+ ['All'],name='measure')
select_radio1 = alt.selection_point(fields=['measure'],bind=bind_radio1)
```

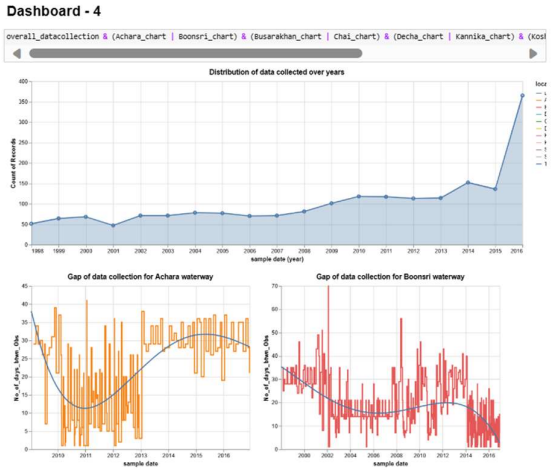
Lead & Zinc chart has two sets of radio buttons. One for selected chemicals and another for selecting the locations. Values are filtered through these interactions and the line chart is shown.

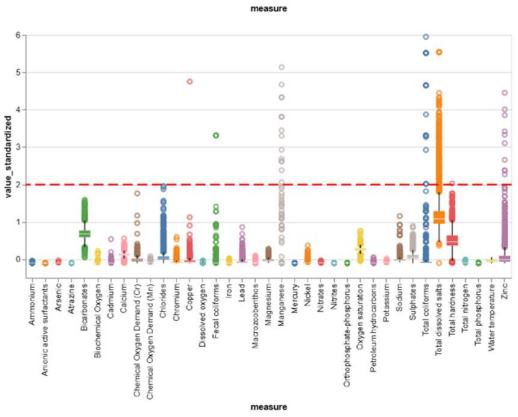
```
#radio button for locations
options_loc1=['Achara','Boonsri','Busarakhan','Chai','Decha','Kannika','Kohsoom','Sakda','Somchair','Tansanee']
labels_loc1 =[option +' '  for option in options_loc1]
bind_radio_loc1 = alt.binding_radio(options=options_loc1+ [None],labels=labels_loc1+ ['All'],name='location')
select_radio_loc1 = alt.selection_point(fields=['location'],bind=bind_radio_loc1)

#radio button for chemicals
options_chemicals1=['Lead','Zinc']
labels_chemicals1 =[option +' '  for option in options_chemicals1]
bind_radio_chemicals1 = alt.binding_radio(options=options_chemicals1+ [None],labels=labels_chemicals1+ ['All'],name='measure')
select_radio_chemicals1= alt.selection_point(fields=['measure'],bind=bind_radio_chemicals1)
```

Conclusion:

These are the analysis on trend and anomalies for chemical contamination. Based on this analysis, we can conclude that **koshoom** is the highly contaminated waterway compared to all other waterways followed by Boonsri and Chai. Koshoom being located close to the dumping site could be the reason for the elevated readings for most chemicals. Strict rules must be followed to stop dumping of harmful waste by Kasios Office Furniture and other industries to stop the decline of RoseCrested Blue Pipit.

DATA QUALITY ANALYSIS





Findings:

- **Distribution of data collection over years chart** – the amount of data collected increased gradually over the years.
- **Gap of data collection frequency chart-** location – Achara, Decha and Tansanee had values only from the year 2009. We could see from the chart that there was an increasing in the data collection in the recent years. 4thorder polynomial trendline is drawn to show the relationship.
- **Boxplot to identify outliers** – we could find some chemicals like iron and total coliforms had a very unrealistic value of 80 standard deviations away from x axis.

Effectiveness of Visual encoding

- **Mark 1,2,3 –** 2-line charts and 1 area chart to show the distribution of data collection over years and gap of data collection frequency. Line charts are used for readability and easy comparison. Area chart to show the spread of data points.
- **Mark 4** - Boxplot to identify the values below the upper and lower quartile range.
- **Mark 5** - Rule to separate the outliers from actual values.
- **Channels** – position (x-sample date, y-value), color- to highlight values based on locations, tooltip to show information when hovered over the datapoint, opacity to show the area.
- **Datatypes -** Temporal: sample date, Nominal: location, Quantitative: value.

Explanation of the Altair code

- Area and line charts are overlapped to create the distribution of data collection over years.
- To find the frequency gap over years – the data is first sorted and grouped as per locations. Then number of days between data collections are calculated as below

```
# Sorting and grouping data
filtered_data=DF.sort_values(by='sample date').groupby(['location','sample date'])[['value']].sum().reset_index()
# calculating the number of days between observation
filtered_data1['No_of_days_btwn_Obs']=filtered_data1['sample date'].diff().dt.total_seconds() / (60 * 60 * 24)
```

- In boxplot, z-score is calculated for standardizing the data and any value which are 2 standard deviations away from x axis are considered outliers. I have also created a rule to highlight outliers.

```
data['value_standardized'] = (data['value'] - mean_value) / std_value
outlier_chart1 = alt.Chart(Filtered_chemicals2).mark_boxplot().encode(alt.X('measure'),alt.Y('value_standardized:Q',scale=alt.Sc
rule_chart1 = alt.layer(alt.Chart(Filtered_chemicals2).mark_rule(strokeDash=[12,6],color='red').encode(y=alt.datum(2)))
```

Conclusion

Data collection must be done frequently and evenly across all locations. Systematic measurement errors should be avoided. Regulatory bodies should be assigned to monitor data collection.