**DATA VISUALIZATION USING ALTAIR**

The analysis begins with the importation of necessary Python libraries, including Altair for visualization, Pandas for data manipulation, NumPy for numerical operations, Datetime for data and time and Isolation Forest for anomaly detection.

The provided dataset is loaded and basic steps like data verification, creation of new columns for year, month, and year\_month, and checking for null values are performed. No null values were detected. Thereafter, we classified the given measures into categories and created a new dataframe (Water\_contamination) which has been used for further analysis.

# Checking for Anomalies:

Anomaly detection is conducted using the Isolation Forest algorithm, and a user-defined function, detect\_anomaly, is created for this purpose. Contamination parameter for anomaly prediction has been taken as 0.1%. The analysis focuses on the "Outliers 0.1%

Contamination" graph, showing the count of anomalies for each measure. Subsequently, anomaly labels are added to the dataframe, distinguishing between Anomaly and Others. Box plots are created for each Anomaly measure with outliers, highlighting locations with error observations. Interactive scatter and bar graphs provide detailed insights into anomaly distribution for specific measures over time.

1. *Interpretation of “Outliers 0.1 % Contamination” graph –*

Measures having taller bars denote higher count of anomalies. Here, the measures have been sorted in descending order of anomaly count of each measure.

Highest anomaly count has been seen in Total dissolved salts, followed by Total coliforms (when considering that the contamination is 0.1%)

1. *Interpretation of Box Plots-*
   1. Anomaly plot: Total dissolved salt – Boonsri has the highest error observation, followed by Sakda, Busarkhan and Decha.
   2. Anomaly plot: Total coliforms – Achara and Kohsoom has highest error observations, followed by Decha and Somchair.
   3. Anomaly plot: Manganese – All locations show some error in observations, except Achara, Decha and Tansanee.
   4. Anomaly plot: Iron – High error in observation are noticed in all locations except Achara, Boonsri, Decha and Tansanee.
   5. Anomaly plot: Zinc – Error in observations can be seen in all locations except Achara and Tansanee.
   6. Anomaly plot: Copper – Significantly high change from the normally observed values is spotted in Kohsoom.
2. *Interpretation of Interactive scatter and bar graph –*
   1. TOTAL DISSOLVED SALTS: Boonsri location exhibited a significant deviation in error observation in August 2006. Notable error observations occurred in the years 2009-2010 and 2013-2016.
   2. TOTAL COLIFORMS: Achara (January 2009) and Kohsoom (October 2010) locations had distinct errors, deviating from the normal range. The years 2009-2013 witnessed several outlier observations, possibly indicating changes in water quality.
   3. MANGANESE: During July and August of 1998-2000 and 2002-2003, Manganese levels showed substantial fluctuations. However, post-2003, all observations fell within the normal range.
   4. IRON: On August 15, 2003, each location experienced one inaccurate observation.
   5. ZINC: The most significant error observation occurred on February 19, 2004, in

Sakda. Observations also exhibited fluctuations in 1999, 2004, 2005, 2010, and 2011.

* 1. COPPER: On August 15, 2003, the Kohsoom region displayed an error observation.

# Trend Analysis:

Trend analysis involves analysing the Water temperature and also creating trend lines for each measure of chemical\_contamination over time, for each location. Interpretations cover various measures such as Arsenic, Chemical Oxygen Demand (Cr), Fertilizer (p, p-DDT),

Total Hardness, Total coliforms and Methylosmoline.

1. *Trend 1-*

We have created a new dataframe focusing on only Water Temperature as the measure and then we have plotted a scatter plot and a line graph for analysis during all the months from year 1998 to 2016.

Interpretation –

The minimum recorded average water temperature is 0.4, while the maximum is 27.1. From January to March and November to December, the water temperature remains below 10 degrees. During June to September, the average water temperature ranges between 20 and 28 degrees. The highest average water temperature was observed in July 2014.

1. *Trend 2-*

We have created an interactive dashboard by plotting trend lines for each measure of chemical\_contamination over the given period of time, for each location.

Interpretation –

* 1. Harmful Metal (Arsenic) : The federal drinking water standard for arsenic set by the Environmental Protection Agency (EPA) is 10 micrograms per liter (µg/L). Prolonged exposure to arsenic in drinking water and food can lead to cancer and skin lesions. Although the values remain below 10 µg/L, with the

exception of Achara and Kohsoom, all other locations exhibit a growing trend in arsenic content in water. Urgent measures are required to address this

increasing trend.

* 1. Dissolved Oxygen (Chemical Oxygen Demand (Cr)) : This is an indicative measure of oxygen likely to be consumed by reactions in a measured solution. So, a higher values of this indicates more severe pollution of organic matter by water. Decha, Kohsoom, Tansanee, and Somchair show a growing trend in dissolved oxygen, exceeding 35. Other locations maintain dissolved oxygen levels between 10-30 by the end of 2016, progressing at a slower pace. Locations with higher value of dissolved oxygen should implement necessary measures.
  2. Fertilizer (p, p-DDT): p, p-DDT is a non-biodegradable pesticide known for causing the extinction of bald eagles due to eggshell thinning. It is banned in most countries. All locations showed some readings of p,p-DDT during 1998- 2001. However, from 2002 onwards, no trace of p,p-DDT has been observed. This indicates no harmful effects from this chemical.
  3. Water Hardness (Total hardness): Total hardness measures the mineral content in water that is irreversible by boiling. Water below 150 mg/L is considered soft, while values over 200 mg/L are deemed hard. Except for Achara and

Decha, all other locations exhibit an increasing trend in total hardness. Busarkhan, Somchair, Kohsoom, and Tansanee have hardness levels exceeding 200 mg/L, showing an upward trend. Necessary actions must be taken to keep values within standard limits.

* 1. Microbial existence (Total coliforms): Total coliform counts provide a general indication of the sanitary condition of a water supply. Due to fewer samples in Achara and Decha, no conclusions can be drawn. However, except for Somchair, all other locations show a decreasing trend in total coliform counts.
  2. Toxicant (Methylosmoline): Methylosmoline is a water-insoluble chemical found in paints, oils, varnishes, dry cleaning liquids, and industrial effluents. It poses risks as carcinogens, reproductive hazards, and neurotoxins, affecting the reproductive capabilities of birds like pipits. Except for Kohsoom and Somchair, all other locations show an average Methylosmoline value declining to zero. Measures should be considered in Kohsoom and Somchair to reduce its presence.

# Data Quality Check:

Data discrepancy analysis includes a heatmap to visualize total samples recorded year-wise for each location. Multiple entries for Iron in 2015 and varying counts of distinct measures highlight data quality concerns.

1. *Checking for missing data and variation in samples recorded -*

A Heatmap has been plotted to see the total number of samples recorded year-wise for each location over the given period of time (1998-2016). This heatmap clearly reveals that locations like Achara, Decha, and Tansanee have data available only from the year 2009. The period from 2005 to 2009 records the highest total observations across all regions. Boonsri consistently made over 1000 total observations every year during 2005-2007. The decline in total observations after 2009 suggests a fluctuation and

decrease in periodic recording of measures across all locations.

1. *Multiple entries for Iron can be seen over different locations in Year 2015 –* Boonsri recorded two entries for Iron in December 2009-2015. Chai had repeated observations on the following dates: February 6, 2015 (3 entries), May 20, 2015 (2 entries), and September 4, 2015 (2 entries). Kannika's entries were repeated on the

dates: January 8, 2015 (3 entries), March 12, 2015 (2 entries), July 2, 2015 (2 entries),

September 24, 2015 (3 entries), and October 8, 2015 (2 entries). Sakda had repeated

observations on ten sample dates in 2015: January 2 (2 entries), February 12 (2

entries), March 12 (2 entries), May 14 (2 entries), July 10 (2 entries), June 18 (2

entries), August 20 (2 entries), October 8 (2 observations), November 19 (2 observations), and December 10 (2 observations). Locations namely Kannika, Sakda and Chai had the highest repeated observations for Iron in 2015.

1. *Checking how many unique measures are present both location-wise and year-wise –* Here click filter has been implemented to interactively select the count of unique measure as per year and also location.
   1. Data quality - There is no regular and consistent recording of unique measures at each location over time.
   2. Location specific - Decha has the lowest count of distinct measures at 42,

while Achara has 45 distinct measures. Tansanee has recorded observations for 55 distinct measures only. In contrast, all other locations have documented more than 90 distinct measures.

* 1. Year specific - Achara, Tansanee, and Decha have recorded observations starting from the year 2009. Between 2005 and 2009, more than 50 distinct measures were observed in all regions, except for the mentioned locations. In the period from 2009 to 2016, there were 50-60 distinct measures observed. Regardless of the region, the highest number of distinct measures was observed in the year 2009.

The analysis concludes with comprehensive insights into anomaly detection, contamination trends, and data quality checks.