Visual Data Analysis Coursework 2

Introduction:

The data given as part of the 2018 VAST Challenge is obtained from a fictitious nature reserve scenario in which a specific bird species, the Rose-Crested Blue Pipit (RCBP), is stated to be endangered owing to the actions of a polluting enterprise. The major goal of the mini challenge is to find proof that either supports or refutes the company's claim that the RCBP is prospering across the preserve. Our technique entails the use of several visual studies to find spatiotemporal patterns.

The following data is supplied with regard to the assignment:
Filename: Boonsong Lekagul Waterways Readings
Description: Primary dataset containing chemical measurements as well as readings collected throughout time at various sites within the preserve.

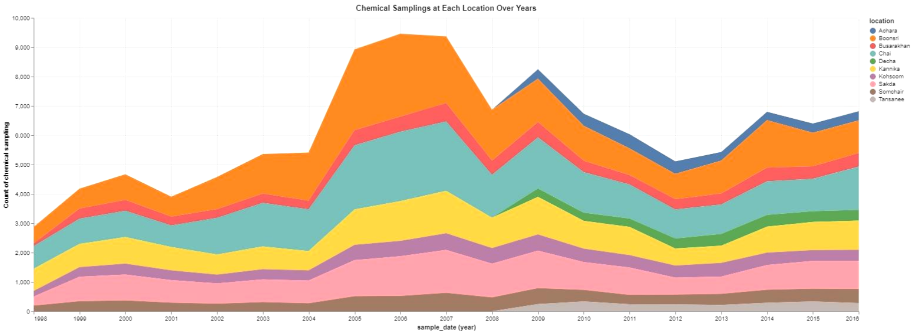
Column Descriptions:
Id: The record's unique identifying number.
Value: The measured value associated with the chemical or attribute in the record.
Location: The name of the location where the sample was collected.
Sample Date: Date on which the sample was collected from the location.
Measure: Chemicals or water properties measured in the record.

Analysis & Visualisation

Finding 1.1

Quality of Finding: The goal of this visualisation is to show the data quality and to highlight if any uncertain issues are present. We focused on finding the missing values in this step.

Dataset Type: Table is the dataset type that has been used here with the values such as sample_date, locations, and the count of the chemical sampling w.r.t location.
Data Type**:** Items and attributes are the data types that are used here.
Attribute Type**:** Quantitative is the type of the attribute type present here.



Actions
Analyse: We have tried to Consume the data and then perform Discover to find out the missing data for various locations over the years.
The visualisation used here to reflect our insight is presented using an **area chart and a heatmap** as proof.

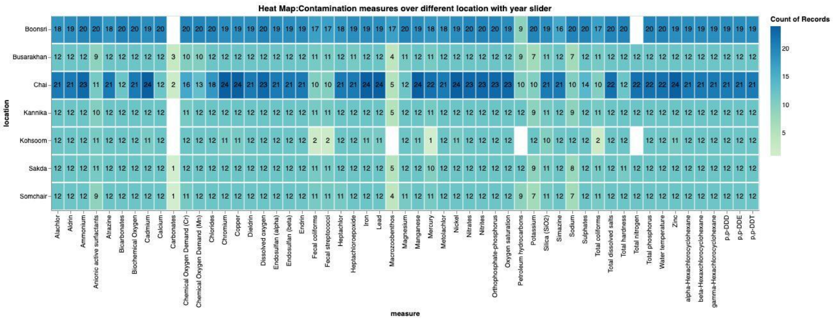
Search: Explore the data to know the count for each measure over a period of time w.r.t location

Query: Identify missing measures for various locations and summarise across all locations.

Target: We discovered a trend in some locations whereby chemical counts increased and decreased in certain years resulting in identifying the trend and the missing values.

Attributes: Distribution of chemical counts in each location over the years

Use of filter/ sorting/ interactions/ dashboard: The filter is used on the year.



Marks: Area chart & Heat map along with the colours

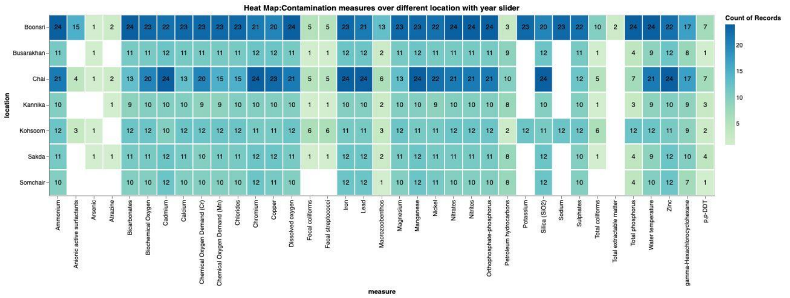
Channels: Position and colour

Position: X-axis - Consists of sample data (year)

Y-axis - Consists of count of chemical sampling

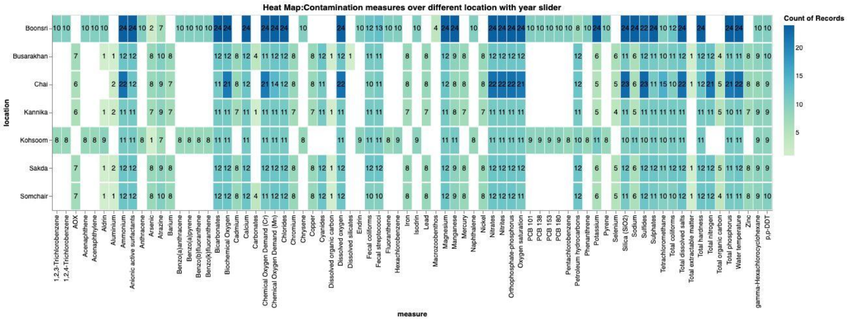
Color Applied to Location for easy identification of locations in the area map in order to find out the highest counts,lowest counts and those with missing values (trend). An area chart was utilised to provide a simple visual view into the varying magnitude of data points in order to spot trends over time.

While in the heat map, colour applied to the count of chemical sampling w.r.t location & years in order to find out the missing values in the location w.r.t time. The darker the colour the more the number of sampling & the lighter the colour, the least number of sampling. With the help of the area chart we were able to find the missing values in the heat map by taking into consideration each chemical count over the location w.r.t years.



Justification for the chosen visual mapping/encoding: The visual channels position and colour were used for the area chart vis. Position which is a magnitude channel for ordered attributes was used for sample_date on the x-axis and count of chemical sampling on the y-axis. Colour is most effective and was used for easy identification of various locations in the area chart. While on the other hand, colour was used for easy identification of the count of the chemical samples in the heat map to visualise the count & the missing values.

Quality of the visualisation and Analysis code: We have tried to find the trend or change in the chemical counts or to find the pattern in the chemical counts based on the locations over a period of time. We were able to find that a few locations such as Boonsri, Chai, Sakda & Kannika were having major contributions to the chemical counts. Also we were able to see the trend that from the year 2004 &



2005 the change in chemical count was rising drastically for a period of time. Also there seems to be no count of samples until the year 2008 in few locations. Henc, we tried to further justify our analysis by means of a heat map. Through heat maps, we were able to see various chemical counts w.r.t location over the years. We were finally able to justify our analysis of finding the missing values in this and saw that from the year 2005 to 2008 there was a huge increase in the chemical counts and also that there were new chemicals being introduced and sampled. Not only this, we were also able to identify a huge number of missing values in the years. It has been observed that:

- Boonsri, Chai, Sakda & Kanika had the sampling from 1998 to 2016.
- There was no sampling for Achara, Decha, & Tansanee from the year 1998 to 2008.
- Kohsoom, Somchair & Busarakhan had missing values which started reflecting more in number after the year 2004.

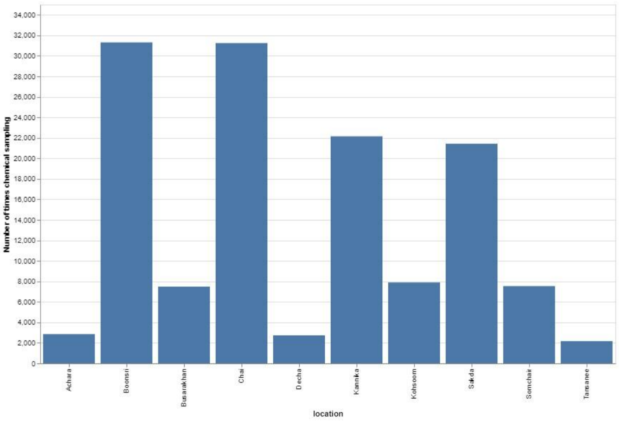
Finding 1.2

Quality of Finding: The goal of this visualisation is to show the data quality and to highlight if any uncertain issues are present. We have focused on finding the change in collection frequency.

Dataset Type: Table is used as a dataset type with the value considered as location, chemical & count of chemical (value).

Data Type: Items and attributes are the types of data type used here.

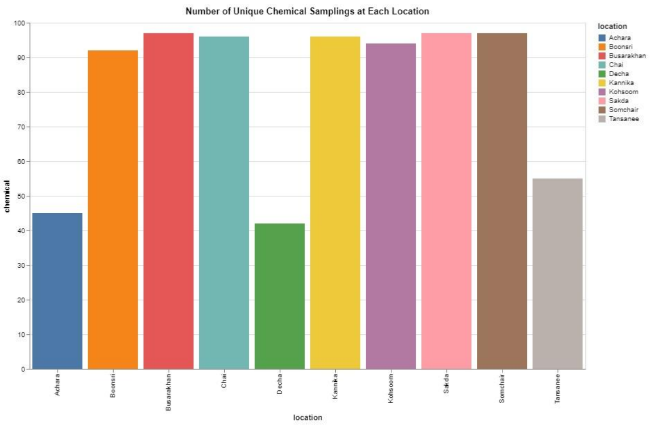
Attribute Type: Quantitative ordered.



Actions

Analyse: We consume the data in order to discover the number of times chemicals were counted at each location so that we could analyse if there was any change in collection frequency. Hence, we have used **Bar charts** to present our findings and new insight was derived from the previous vis as the number of unique chemical samplings at each location.

Search: To know how many times each chemical was counted over the years, locate was used to check through every location and to verify the change of frequency w.r.t location.



Query: Summarise the counts of chemicals from each location then deriving unique chemicals samplings from them.

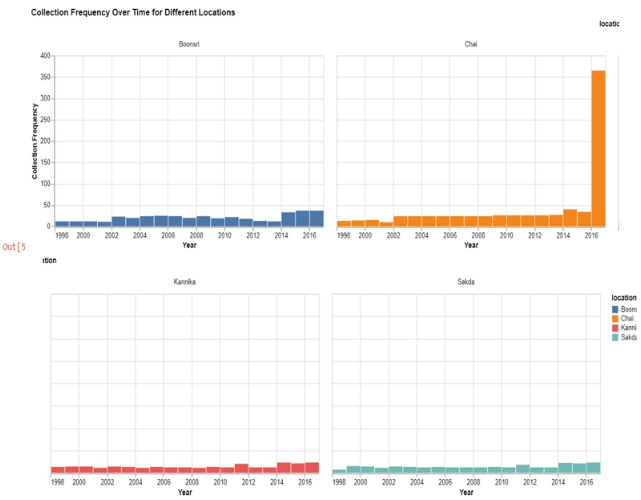
Target: We discovered a trend in some locations whereby chemical counts are more in number while in some locations they are in less number resulting in identifying the trend. Further we tried to maximise our insight by targeting unique chemical count & later to target specific locations based on the inputs.

<u>Attributes</u>: Distribution of the chemical/ sampling in different locations.

Marks: Line was used for all visualisations
Channels: Position and colour is used as a channel.
Position: X-axis - Consists of location and year
Y-axis - Chemical sampling and collection of frequency
Color Applied to locations in the number of unique chemical sampling at each location.

Use of filter/ sorting/ interactions/ dashboard: No filter/ sorting is used in this visualisation.

Justification for the chosen chart type: We wanted to get insight in terms of finding the change in collection frequency from the given dataset. From our above analysis, we tried to further analyse it by using a barchart that was highly effective for this insight to allow a viewer see at once what location or chemicals were the most or least counted.



Justification for the chosen visual mapping/encoding: The position and colour of the visual channels were employed for the viz. Position, a magnitude channel for ordered qualities, was utilised for both location and year on the x-axis, as well as chemical sampling and frequency collection on the y-axis. Colour is the most efficient and was used to easily identify distinct locations.

Quality of the visualisation and Analysis code: We have tried to analyse and visualise the change in collection frequency in different locations with the help of the bar chart. Here, the bar chart is helping in a very informative way to get the right insight for our consideration. We first tried to find the total count of sampling of the chemicals across the locations which helped us in understanding which location has the major contribution of the chemicals. Further, we tried to see by means of unique chemical count which locations are having the highest count. By this we were able to justify which locations were actually having the maximum values and we further tried to get insights from those locations.
In this scenario, we considered the four locations - Boonsri, Chai, Kanika & Sakda; as they had the maximum count of the samples. When we further tried to find the change in frequency we can see from the visual from the above that the **Chai** had a sudden change in frequency with respect to the year 2016. Hence, we were able to find change in collection frequency w.r.t location and chemical over a period of time.

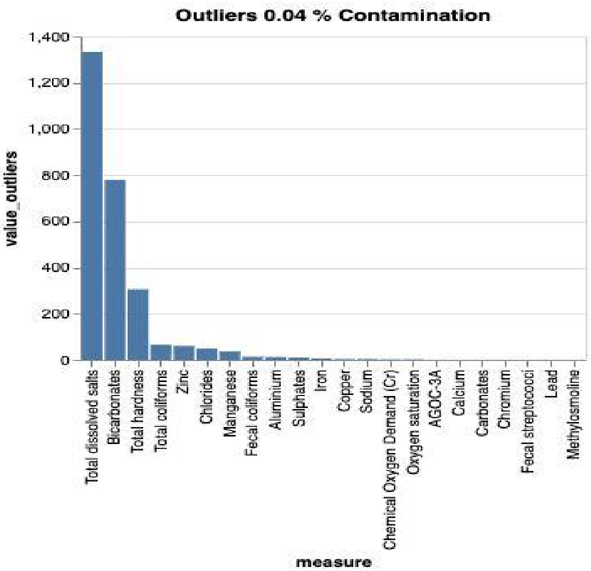
Finding 1.3
Quality of Finding: The goal of this visualisation is to show the data quality and to highlight if any uncertain issues are present. We have focused on finding the unrealistic values in the present scenario..

Dataset Type: Table is used as a dataset type with the value considered as chemical (measure) & count of chemical (value).

Data Type: Items and attributes
Attribute Type: Quantitative ordered attribute.

Actions

Analyse: In this, we are trying to analyse the unrealistic value present in the data to retrieve the outlier. For this it is considered as Discover outliers.



Search: To identify or find out the unrealistic value without knowing the chemical type and the location where it can be present, Explore is the type of search feature used here.
Query: Identify, in order to get the required insight for the unrealistic values from the dataset.
Target: Outliers is our target as they can be represented very well to find the unrealistic value.
Attributes: Distribution and under that Extremes

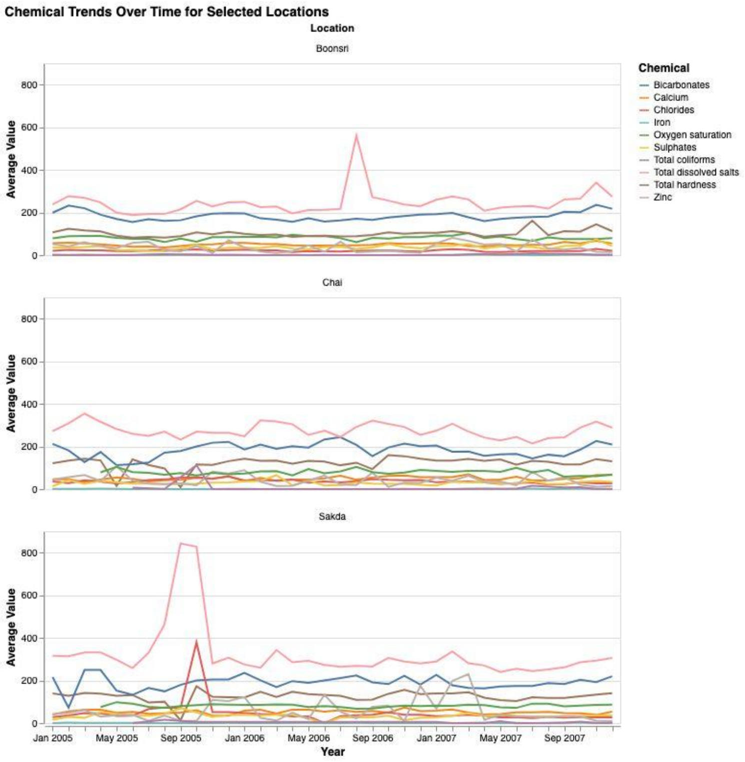
Marks: Line was used for this visualisation
Channels: Position is used as the channel here
Position: X-axis - Consists of Measures
Y-axis - Consists of Value_outliers
Use of filter/ sorting/ interactions/ dashboard: No such filter/ sorting has been used for this visualisation.

Justification for the chosen chart type: In order for us to get insight in terms of finding unrealistic values in the dataset, barchart proved to be highly effective for this insight to allow a viewer see at a glance what chemical has the highest amount of outliers compared to other chemicals in the dataset.

Justification for the chosen visual mapping/encoding: The position of the visual channels was used for the viz. Position, a magnitude channel for ordered qualities, was utilised for measures on the x-axis, and value_outliers on the y-axis.

Quality of the visualisation and Analysis code: We have tried to analyse and visualise the unrealistic value that might be present in the dataset which was unknown by means of chemical and the location. Here, the bar chart is helping in a very informative way to get the right insight for our consideration. With the help of the above visuals we were able to get the unique count and sampling count of the chemicals. By which further we tried to see and analyse which of the chemicals are acting as unrealistic values. By this we were able to justify which chemical was actually having the unexpected sudden difference in its count/values.
When we further tried to find the unrealistic value (outlier) we can see from the visual from the above that the **Total dissolved salts, bicarbonate & total hardness** had a sudden change in frequency and hence, has the highest peak in the graph highlighting it as an unrealistic value. If this total dissolved salts & bicarbonate are present in maximum amount in the water which will not only harm the aquatic life but yes can be dangerous to the birds.

Finding 2.1



Quality of Finding: The purpose of this visualisation is to find the trend over a period of time either with respect to the location of the chemicals. And we decided to find the trend w.r.t selected locations with the top impactful chemicals.
Dataset Type: Table is a dataset type with the value in consideration as chemicals, location, year & value.
Data Type: Items and attributes
Attribute Type: Quantitative ordered

Actions
Analyse: In this visual, we have tried to Consume the data from the dataset in order to discover the trend over a period of time either at a location or with the chemicals to present the result as in which can help us identify the trend of it.
Search: To identify or find out the trend of the given data by means of selected location with the top most effective chemicals over a period of time. Where Browse is the type of the search feature incorporated here.

Query: Compare is the type query used here to find the trend by means of comparing the top most effective chemicals w.r.t the selected locations.
Target: Trends is our target in order to visualise how over a period of time the chemicals have impacted the selected locations.
Attributes: Here there is a multiple attribute which is a type of dependency as the value of the selected chemicals are forming a trend w.r.t the time.
Marks: Line was used for this visualisation
Channels: Position and colour are used as channels for this visualisation
Position: X-axis - Consists of Year
Y-axis - Consists of Average value
Color Applied to Chemicals.
Use of filter/ sorting/ interactions/ dashboard: We have selected or filtered the chemicals by means of the top most effective one. Aso, filter is applied to the location for the only selected locations.

Justification for the chosen chart type: For this visualisation, we have used a line chart because it allowed us to compare trends over time across selected locations & the top chemical which has the major contribution in the sampling values.. When compared to the others, this chart type proved to be the most effective.

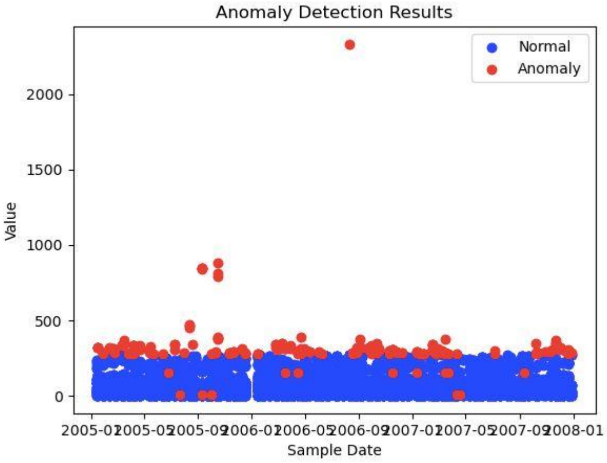
Justification for the chosen visual mapping/encoding: The position and colour of the visual channels were employed for the visualisation. Position, a magnitude channel was used for years on the x-axis and average value on the y-axis. Colour is helpful in identifying selected chemicals used in the dataset.

Quality of the visualisation and Analysis code: We have tried to analyse and visualise the trend over the year by considering the top most impactive chemicals such as bicarbonates, total dissolved salts, total hardness chlorides, etc in the selected locations which had the more chemical contribution such as Boonsri, Chai, & Sakda. Here, the line chart is helping in a very informative way to get the right insight for our consideration. With the help of the above visuals we were able to get the trend for the top chemicals over a period of time in the selected locations.
With the help of the visualisation we can see that the **Total dissolved salts, bicarbonate & total hardness** had a huge contribution or the highest contribution, also it has the highest peak in the graph w.r.t other chemicals. These chemicals can harm the living species if found in major value in the water.

Finding 2.2
Quality of Finding: The purpose of this visualisation is to find the anomalies present in the dataset over a period of time, & location with respect to the chemical pollution.

Dataset Type: Table is a dataset type with the value in consideration as chemicals, location, year & value.
Data Type: Items and attributes
Attribute Type: Quantitative

Actions
Analyse: In this visual, we have tried to Consume the data from the dataset in order to discover the anomalies over a period of time at a location with the chemicals pollutants to present the result as in which can help us identify the major anomalies present in it.
Search: To identify or find out the anomalies value without knowing the chemical type at the selected location over a period where it can be present, Explore is the type of search feature used here.

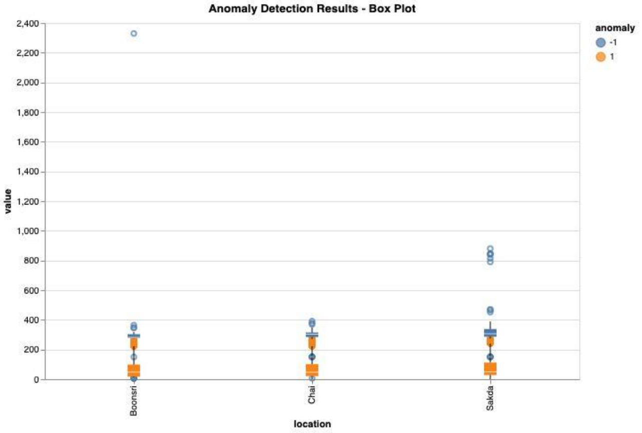


Query: Identify is the type of the query used here in order to find out the anomalies present in the chemicals by targeting the selected locations which had a major contribution of the sampling values.
Target: Outliers is our target as they can be represented very well to find the anomalies value.
Attributes: Distribution is the attribute type used here under which falls the extremes.

Marks: Scatter chart & Box plot was used here to present our visual & insight.
Channels: Position and colour
Position: X-axis - Sample date (fig 1) Location (fig 2)
Y-axis - Value
Color Applied to the filter; normal and anomaly

Use of filter/ sorting/ interactions/ dashboard: Filter is implied on the location.

Justification for the chosen chart type: A scatter plot and a box plot were used to monitor the relationships between all chemicals plotted, also it allows us to check the distribution of the data points, which can help identify anomalies in our dataset. Other chart types, such as bar charts and line charts are not well-suited. While box plot significantly helps in identifying the anomalies.



Justification for the chosen visual mapping/encoding: The position and colour of the visual channels were introduced for the visualisation. Position, a magnitude channel was used for sample_date and locations on the x-axis and value on the y-axis. Colour is helpful in identifying the anomalies and the chemicals that are normal in the dataset.

Quality of the visualisation and Analysis code: We have tried to analyse and visualise the anomalies present at the selected location over the year by considering the top most impactive chemicals such as bicarbonates, total dissolved salts, total hardness, chlorides, etc (selected locations with maximum contribution of the chemical samplings such as Boonsri, Chai, & Sakda). Here, the scatter chart in the first visual is helping in identifying the anomalies in a very informative way to get the right insight for our consideration by highlighting the colour red as anomalies and blue as normal.
When we further try to find the exact anomaly type (outlier) we can see from the second visual (box plot) that the **Total dissolved salts, bicarbonate & total hardness** falls as anomalies in the selected locations (Boonsri, Chai & Sakda) in huge counts. If this total dissolved salts & bicarbonate are actually present in maximum amount in the water it will not only harm the aquatic life but yes can be dangerous to the birds.

Conclusion:

The data analysis and visualisation reveal the trends of chemical pollution in the Boonsong Lekagul Wildlife Preserve over time. While the broad patterns in chemical concentration are clear, the occurrence of missing data casts doubt on the chemical levels during certain times.

The abnormalities discovered in several areas imply that chemical contaminants from a neighbouring industry may have entered the preserve's rivers and streams, causing rapid increases in concentration measurements. However, flaws in data gathering, such as missing records and uneven sample rates, call into question whether these spikes are truly significant, especially because the specific dates of the increases are unclear.

Additionally, the Blue Pipit birds in Boonsri, Chai and Sakda are under threat, owing to rising levels of the total dissolved salts, Bicarbonates & Chlorides in certain years. Other chemical pollutants may potentially endanger aquatic life and birds in a variety of ways. Implementing a more rigorous sampling technique and adding more information on the bird population might improve data processing and visualisation, offering greater clarity into the challenges at hand.

Team Members:

Alisha Vaibhavi Adhav	**M00955828**
Miracle Blessing Paulinus	**M00799831**
Micheal Frimpong Dankwah	**M00953035**
Muhammad Khurram Shahzad	**M00972373**
Junaid Dawood Momin	**M00904599**