MSc Data Science
CST4060 Visual Data Analysis
Data Visualization Using Tableau

By: Hinal Gala – M00905663

Ahalya Marchala – M00920247

Eleman Zaiko – M00886253

Tushar Raj – M00915939

Angad Partap Singh – M00912257

INTRODUCTION

The data we are considering is the collation of sensor readings from Boonsong Leekagul Waterways.

These readings are of different chemicals at various data detection points.

The readings are taken over a period of 18 years (1998 – 2016).

Using this data, we will be able to understand the contamination levels at different sites. We would look at the trends, some anomalies, fluctuations etc.

RESEARCH QUESTIONS

1. Describe any data quality and uncertain issues, such as

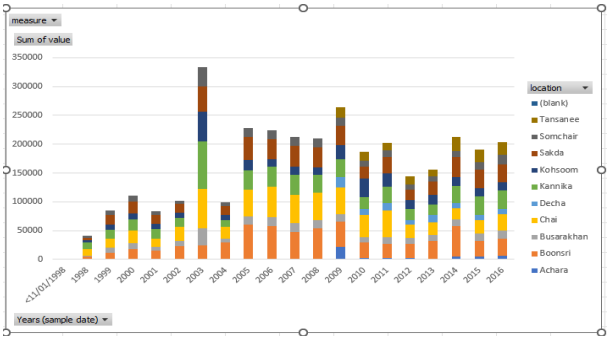
- missing data
- change in collection frequency.
- unrealistic values (e.g., water temperature higher than 100 degrees).

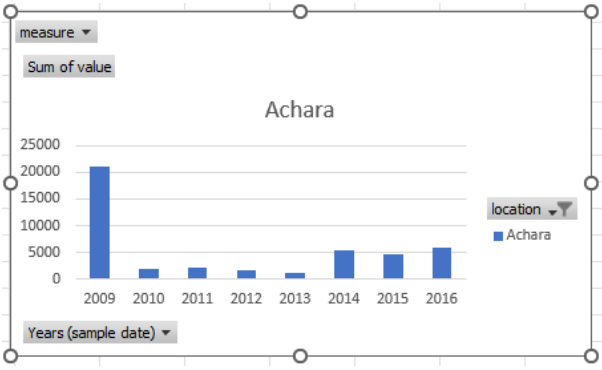
2. Describe trends and anomalies with respect to chemical contamination

- Trends: changes over time and/or sensor site
- Anomalies: sudden change over time or one site significantly different from others

Question – 1

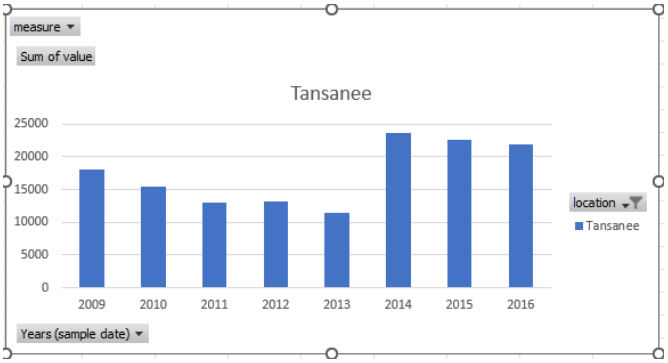
Missing Data





We tried to analyze how different locations gave readings over the period over time.

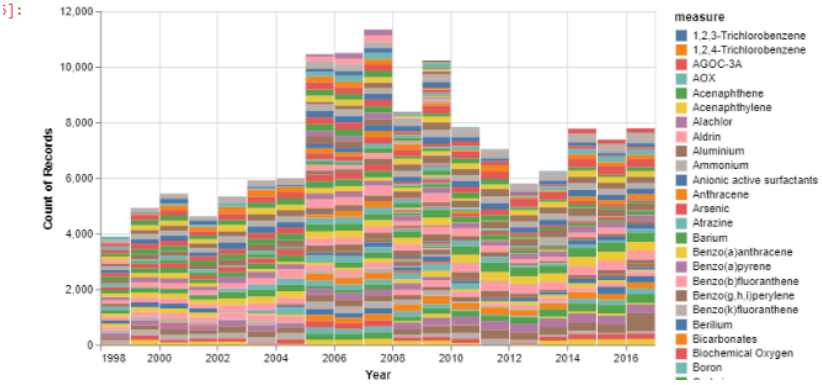
- The graph represents each location recorded certain contamination level apart from ACHARA until 2009. Later the values dipped down significantly.
- Another location “TANSANEE” didn’t show any values till 2009. The reason behind this could be that the sensors wouldn’t have been installed until 2009.



This situation can be inferred as MISSING VALUES as no reading from 1998-2009. However, the contaminations were there already.

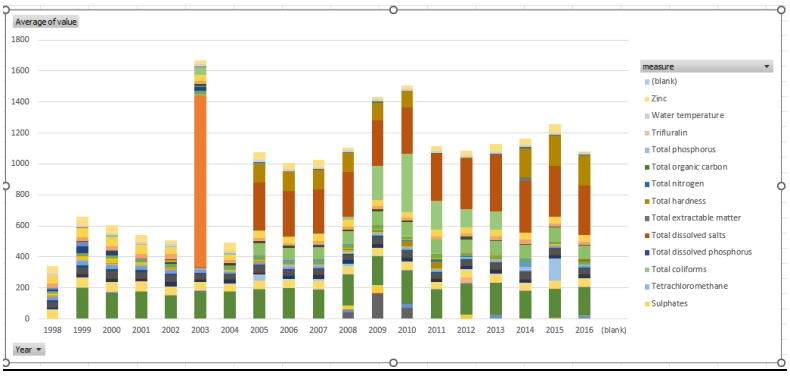
Change in Collection Frequency

```
alt.Chart(df).mark_bar().encode(
    x=alt.X('year(sample date):T',title='Year'),
    y='count(distinct(measure)):Q',
    color='measure:N'
).properties(
    width=500
)
```



In the year 2007, there were some new sensors installed because of which we can see the increase in frequent readings.

Unrealistic Values



In the year 2003, the orange point (Iron), we see a significant jump in the frequency of Iron. The recording that we received, the maximum average of readings that we received was from Iron compound. In comparison to rest of the years, we see an UNREALISTIC CHANGE in readings that year.

QUESTION-2

Trends: changes over time and/or sensor site

Finding 1

Insight 1: Overall trend of contamination levels over time at highest contaminated location

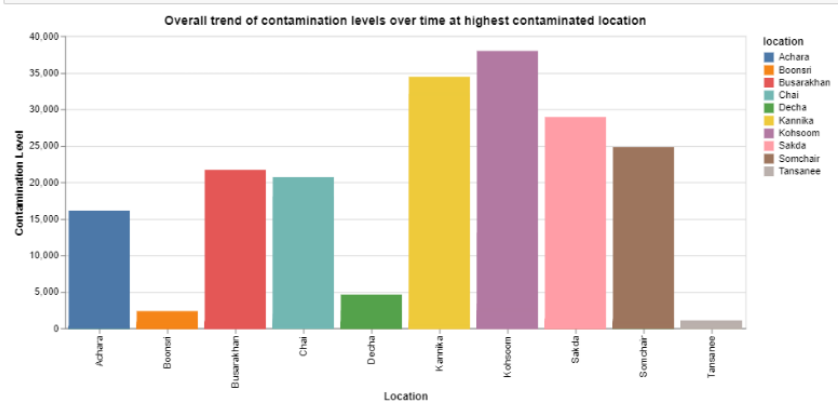
We have used bar chart as they effectively depict any trend changes over time and area chart are helpful in showing rise and fall of data.

We are trying to find the location with highest contamination.

```
chart1 = alt.Chart(df).mark_bar().encode(
    x=alt.X('location:N', title='Location'),
    y=alt.Y('value:Q', title='Contamination Level'),
    color='location:N',
    tooltip=['location', 'value']
).properties(
    width=700,
    title='Overall trend of contamination levels over time at highest contaminated location'
)

chart1
```

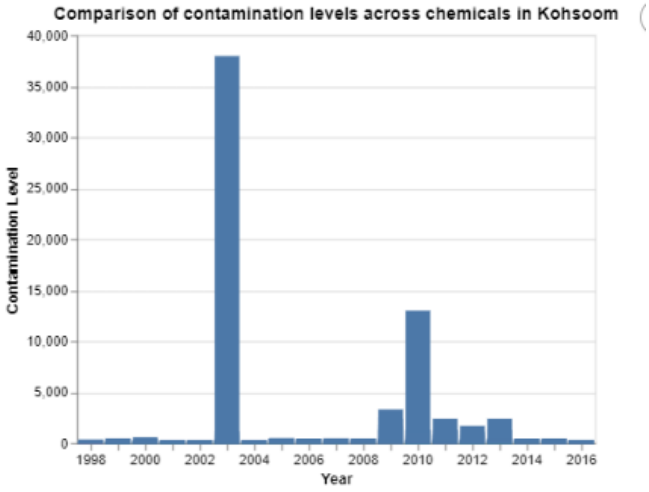
The result shows that Kohsoom is the most contaminated location of them all.



Insight 2: Comparison of contamination levels between different chemicals to find highest value for chemical in Kohsoom.

As we have discovered Kohsoom is the highest contaminated location. To dig deeper, we wish to find which chemical holds the maximum value. From the results we can see that in 2003, Iron holds the highest contamination value.

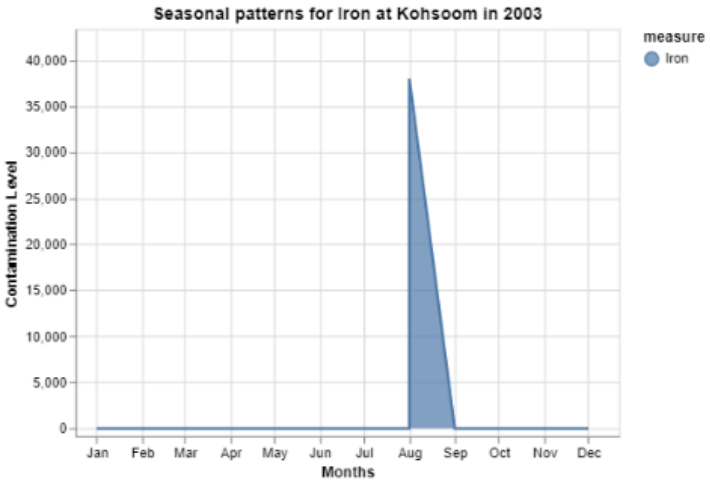
```
chart2 = alt.Chart(new_df).mark_bar().encode(
    x=alt.X('year(sample date):N', title='Year'),
    y=alt.Y('value:Q', title='Contamination Level')
).properties(
    title='Comparison of contamination levels across chemicals in Kohsoom'
)
```



Insight 3: Seasonal patterns for Iron at Kohsoom in 2003

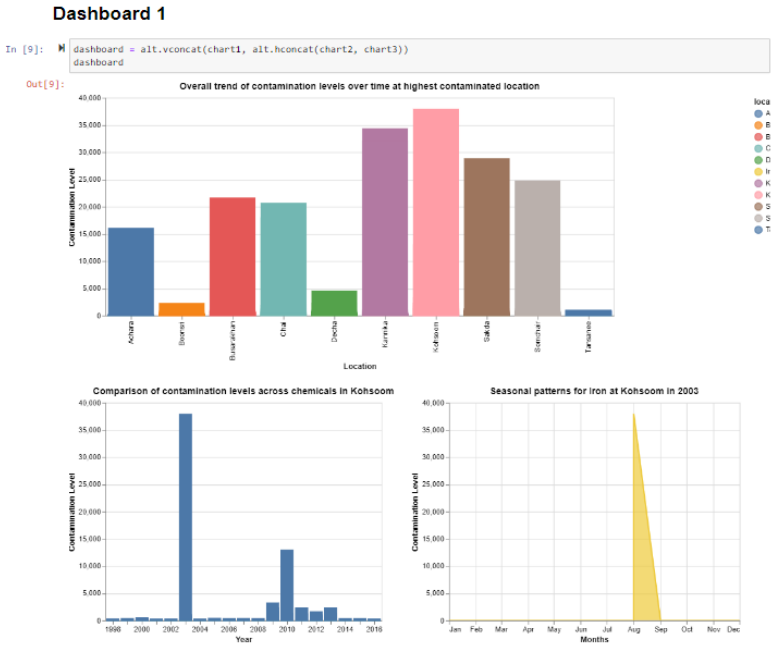
This chart represents the trend in Iron contamination value for the year 2003. There is the sudden change in the values in the month of August. The value went up to 37,959.28.

```
# Create an area chart with interactive selection
chart3 = alt.Chart(new_df_filtered).mark_area(
    line={'color': 'white'}
).encode(
    x=alt.X('month(sample date):T', title='Months'),
    y=alt.Y('value:Q', title='Contamination Level'),
    color='measure:N',
    tooltip=['measure', 'value']
).properties(
    title='Seasonal patterns for Iron at Kohsoom in 2003'
).interactive()
```



DASHBOARD 1

The Dashboard shows the interaction between the graphs in Finding 1.



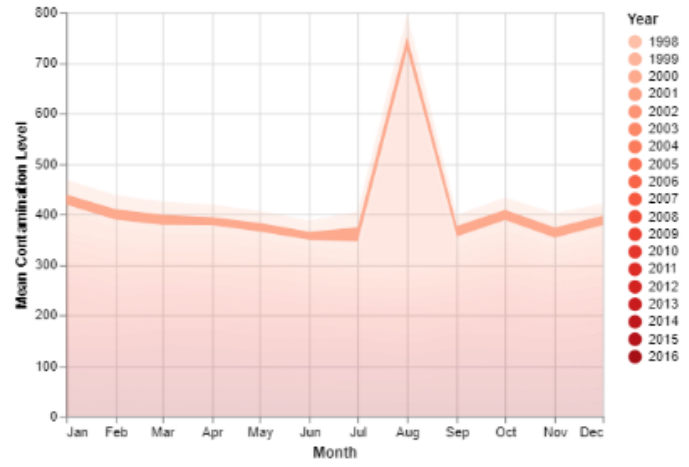
Finding 2

We have used area plot to show denser part of the year. Also used colour, and the drop-down menu for the year-selection. Further we've used scatterplot to show distribution of the measure levels. Also tool-tip t display the additional info. The bar-graph helpful for min and max chemical values.

Insight 1: Comparison of chemical contamination levels across different sensor sites at the same point in time

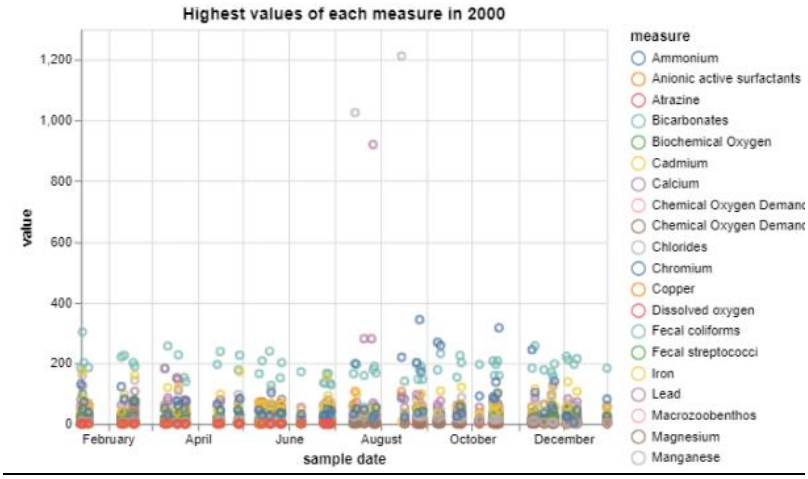
Here we are comparing the levels of contamination at all the sites. We are taking into consideration only one timeframe (for eg- readings in the year 2000).

```
chart1 = alt.Chart(df_year).mark_area(opacity=0.7).encode(
    x=alt.X('month(sample date):T', title='Month'),
    y=alt.Y('mean(value):Q', title='Mean Contamination Level'),
    color=alt.Color('Year:N', title='Year', scale=alt.Scale(scheme='reds')),
    tooltip=['Year', 'mean(value)', 'month(sample date)'],
    opacity=alt.condition(year_selection, alt.value(1), alt.value(0.2))
).add_selection(year_selection)
```



Insight 2: Highest value of each measure in 2000

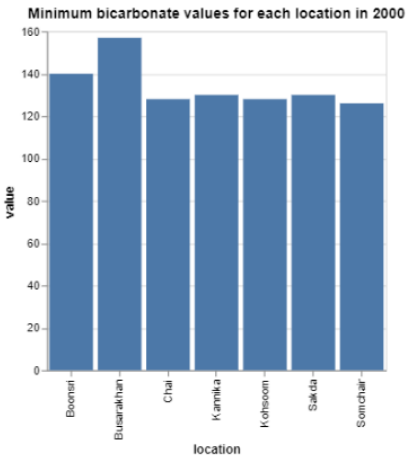
```
chart2 = alt.Chart(highest_values).mark_point().encode(
    x='sample date:T',
    y='value:Q',
    color='measure:N',
    tooltip=['location', 'measure', 'value', 'sample date']
).properties(title='Highest values of each measure in 2000')
```



This plot shows the highest contamination level per each measure in the year 2000. We found that BICARBONATES are having the constant values in the year 2000 across all the locations.

Insight 3: Minimum Bicarbonate values for location in 2000

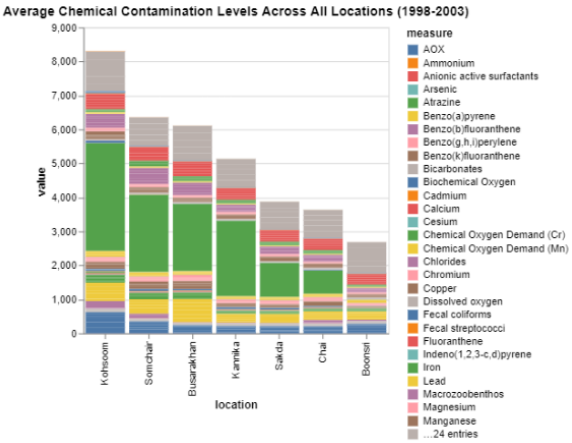
```
chart3 = alt.Chart(lowest_bicarb).mark_bar().encode(
    x='location:N',
    y='value:Q',
    tooltip=['location', 'value']
).properties(width=300, title='Minimum bicarbonate values for each location in 2000')
```



We found that all the locations had approximately same levels of BICARBONATE contamination in the year 2000.

Additional Statistical Analysis

```
chart4 = alt.Chart(avg_data).mark_bar().encode(
    x=alt.X('location', sort=alt.EncodingSortField(field='value', op='mean', order='descending')),
    y='value',
    tooltip=['measure', 'location', 'year', 'value'],
    color='measure'
).properties(
    title="Average Chemical Contamination Levels Across All Locations (1998-2003)",
    width=300
```

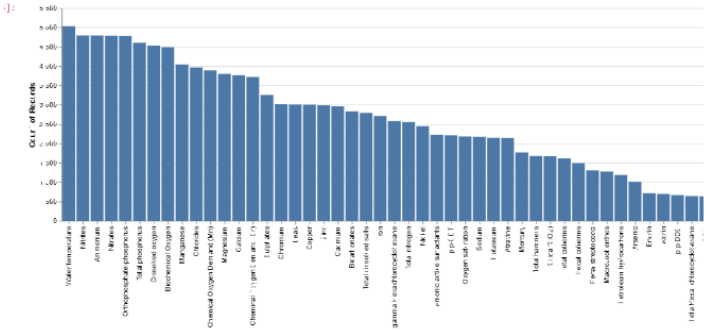


Dashboard 2

FINDING 3

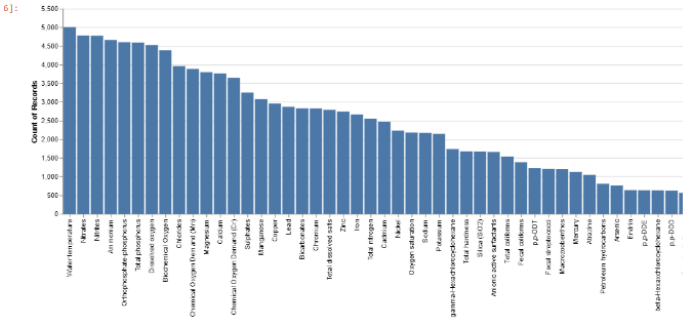
We've used Mark= Bar and Channel= Sort. This has been helpful to find observations for a period. We've used scatter plot as well, so as to see the outliers.

```
alt.Chart(df).mark_bar().encode(
    x=alt.X('measure:N', sort='-y'),
    y='count(measure):Q'
)
```



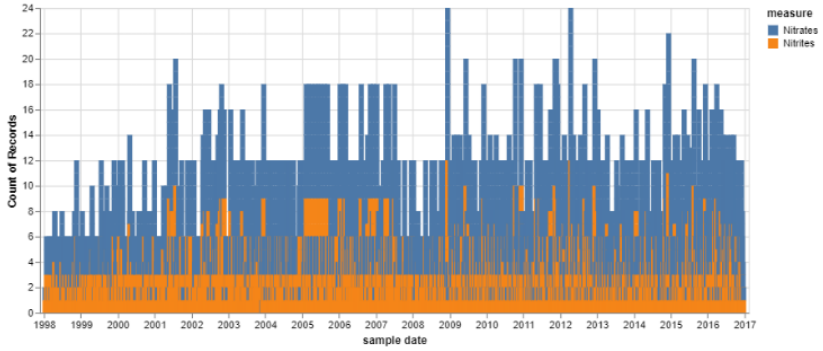
This finding is about understanding the frequency distribution of chemicals. We've chosen bar-graph as we aim to plot the frequency distribution. We've observed here that Nitrides and Nitrates ammonium, orthophosphate hyper phosphorus are equally distributed across the sites.

```
chart_1 = alt.Chart(df).mark_bar().encode(
    x=alt.X('measure:N', sort='-y'),
    y='count(measure):Q'
).transform_filter(
    datum.value != 0
)
chart_1
```



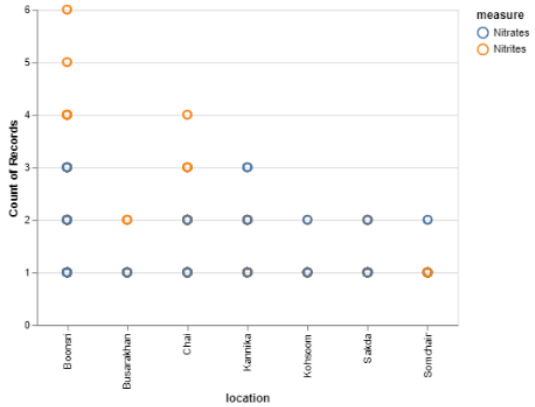
This finding is about understanding the frequency distribution of chemicals whose sensor readings are not Zero. We've chosen bar-graph as we aim to plot the frequency distribution. We've observed here that Nitrites and Nitrates are equally distributed across the sites. There is a dip in the other two chemicals (ammonium &orthophosphate hyper phosphorus). This tells us there were readings of these two chemicals as zero. We've used Mark= Bar and Channel= Sort

```
chart_2 = alt.Chart(df).mark_bar().encode(
    x="sample date:T",
    y="count(measure):Q",
    color='measure',
    tooltip=['location', 'value', 'sample date']
).transform_filter(
     alt.FieldOneOfPredicate(field='measure', oneOf=['Nitrites', 'Nitrates'])
).properties(
    width=700
)
```



Now, we have realized that Nitrites and Nitrates are only two chemicals that have been dumped the most into the water and adding the contamination levels. Therefore, we have plotted the graph of these two chemicals across the years to understand how they are distributed. We observed that in 2008 and 2012, there has been sudden spike in the dumping of these chemicals. We have used filtering to filter the two chemicals.

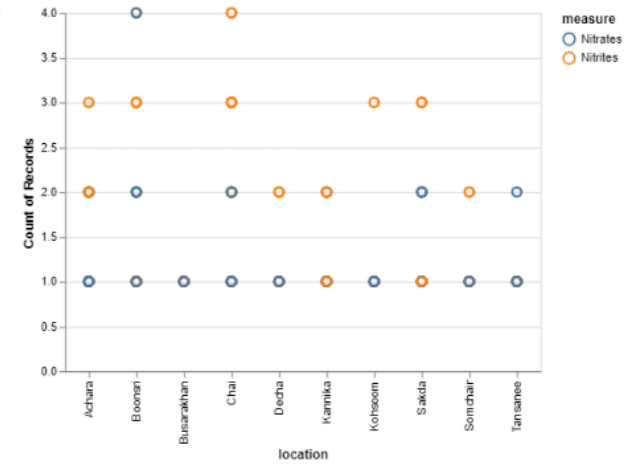
```
chart_3 = alt.Chart(df_2008).mark_point(size=60).encode(
    x="location:N",
    y="count(measure):Q",
    color='measure',
    tooltip=['measure', 'value']
).transform_filter(
     alt.FieldOneOfPredicate(field='measure', oneOf=['Nitrites', 'Nitrates']),
).properties(
    width=400
```



This finding is about understanding the rate of dumping of Nitrites and Nitrates across the locations in the year 2008. For site Sakda and Koshoom, they don't have Nitrites sensor, and Boonsri location is highly contaminated with these two chemicals. We've used filtering to filter over the chemicals.

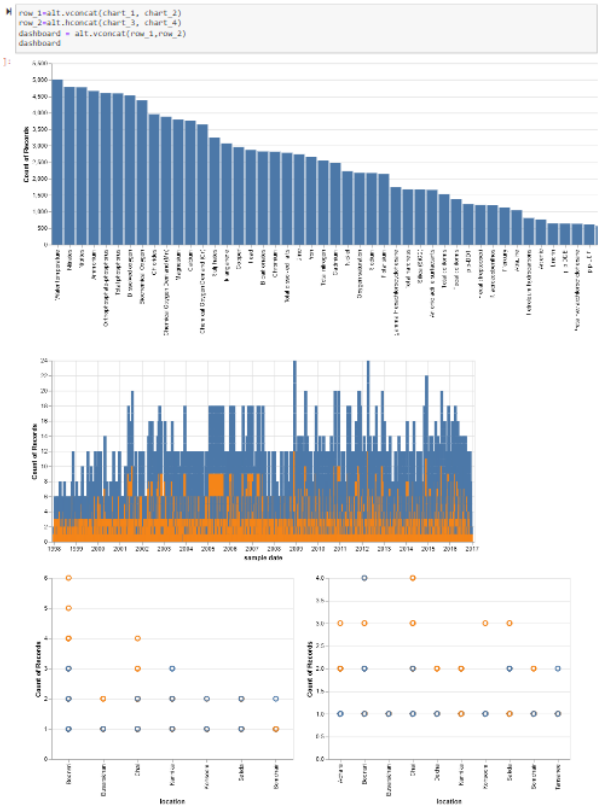
The scatter plot is important here because it helps us identify the levels of distribution of these chemicals.

```
chart_4 = alt.Chart(df_2012).mark_point(size=60).encode(
    x="location:N",
    y="count(measure):Q",
    color='measure',
    tooltip=['measure', 'value']
).transform_filter(
     alt.FieldOneOfPredicate(field='measure', oneOf=['Nitrites', 'Nitrates']),
).properties(
    width=400
```



This finding is about understanding the rate of dumping of Nitrites and Nitrates across the locations in the year 2012. We see some additional sites like Tansanee, Decha and achara. This means that new sensors have been installed at these locations. Tansanee doesn't have Nitrites sensor, and Boonsri and Chai locations are highly contaminated with these two chemicals. We've used filtering to filter over the chemicals.

Dashboard -3



Conclusion :

Anomalies:

- No data of Achara before 2008, and a sudden spike only in 2009, comparing to other years.
- Sudden increase of Iron in August (2003), and sudden decline in the later years.
- New readings from Newly installed sensors in Tansanee, Decha and Achara were found.

Trends:

- Koshoom being the only location which was highly contaminated across all the years.
- Nitrites/Nitrates have been found the most across the sites. This leads to the possible contamination in the water as they are highly reactive.