Data Visualization with Altair

Analysis Framework

The data that was given to analyze was Attributes and Items. There were values given for certain attributes and items.

The dataset type was tables. Every attribute and item and its values were arranged in a table.

Some attribute type was quantitative and some of it was categorical.

From the data that were given which were sample date, id, measure, location, and value. I thought about making a visualization showing how the data is changing over the years and across different locations. But before doing that I thought about checking which data I can use to show it. I had to check which data will be useful for showing the visualization I want.

The actions that I did is shown below:

Analyze: I analyzed the data that was given to me. I thought about showing the data that was presented to me and discovering a pattern from it.

Search: I searched for the data which I'll need to use for my task. I had to eliminate the data which won't be useful. So I had to explore the data.

Query: After that, I thought about using the data to compare. Comparing data over the years and for all locations.

Targets: First my target is the measures and check whether there were any missing values over the years abnormal values. Which will hamper me from making a pattern. After that, I can get the measures by which I can make a pattern of.

Visual-1:

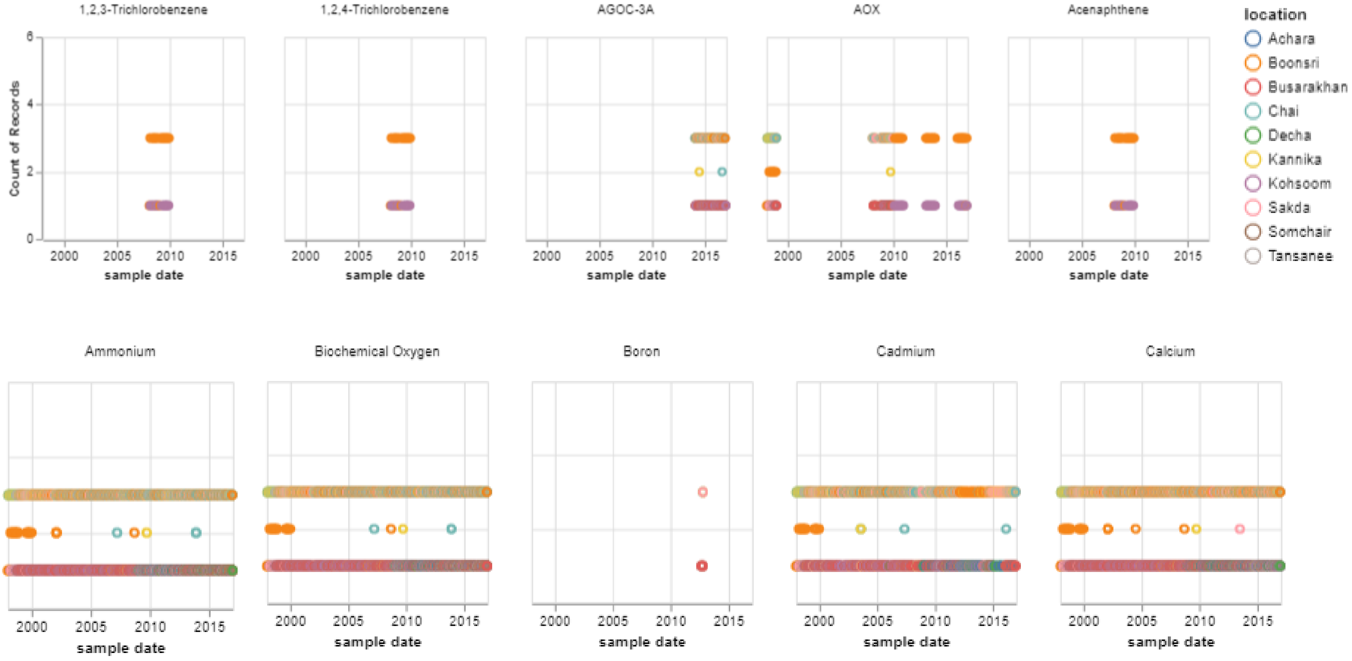
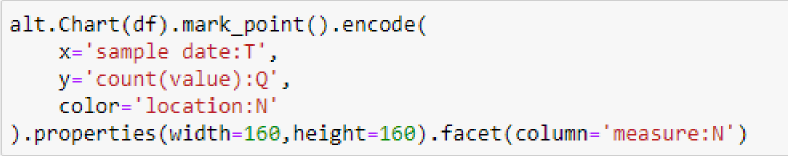


Figure 1



```
alt.Chart(df).mark_point().encode(
    x='sample date:T',
    y='count(value):Q',
    color='location:N'
).properties(width=160,height=160).facet(column='measure:N')
```

For the first visualization, I showed the count of values over time for each measure to find out the data collection frequency. I used the quantitative attribute to y coordinate and the categorical attribute to color. I used a zero-dimensional mark: point for it. To identify different marks, I used the color channel. I think for showing a data frequency over time to check if there are any multiple entries or any missing values this is an appropriate visualization. I have just shown a few of them in the ‘fig-1’.

From this visual, we can see that there were multiple entries of the same data. So, the collection frequency was very much unusual. There were also missing values for many measures over time.

Visual-2:

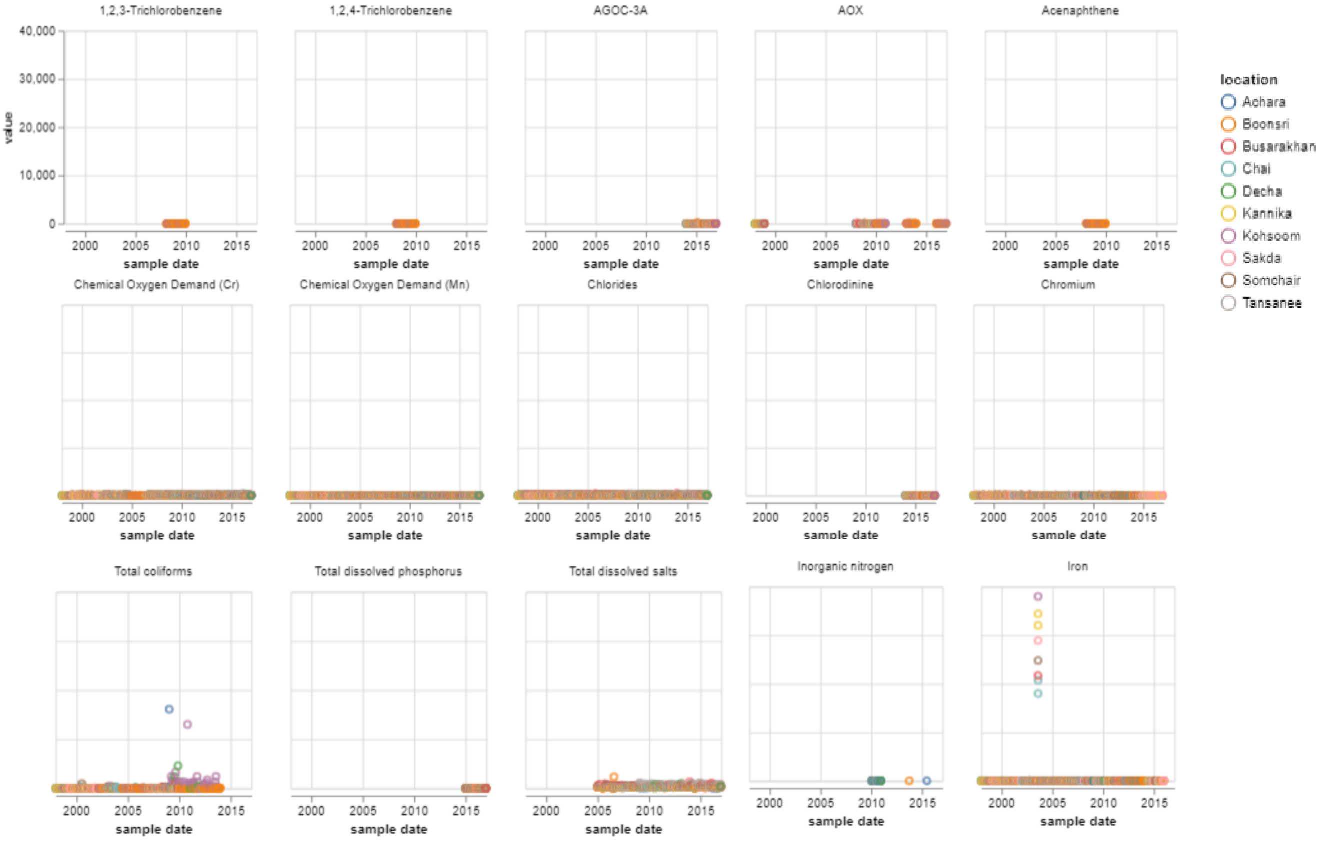
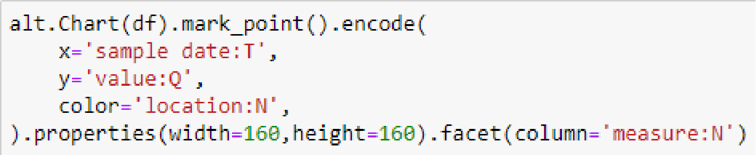


Figure 2



```
alt.Chart(df).mark_point().encode(
    x='sample date:T',
    y='value:Q',
    color='location:N',
).properties(width=160,height=160).facet(column='measure:N')
```

For this visualization, I showed the values over time for each measure to find out the abnormal data among them. I used the quantitative attribute to y coordinate and the categorical attribute to color. I used zero-dimensional mark: point for it. To identify different marks I used the color channel. I think for finding an abnormal data this visualization would be effective to separate those outliers. I have shown just a few of them in ‘fig-2’.

From the visualization, we found some measures with unusually high values.

Visual-3:

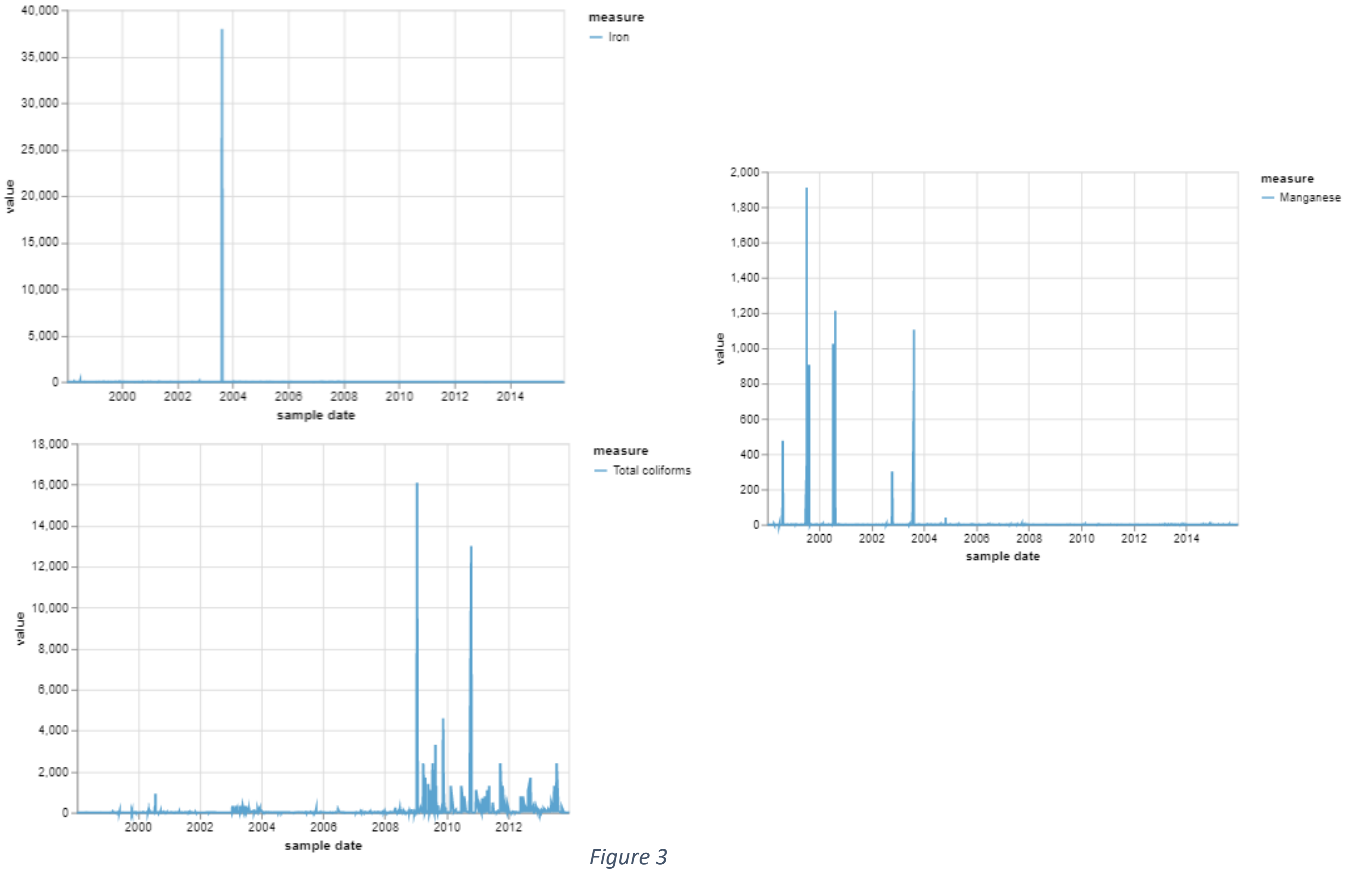
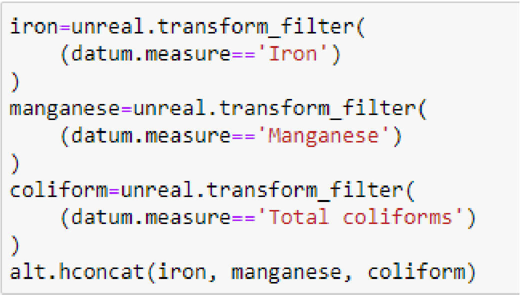
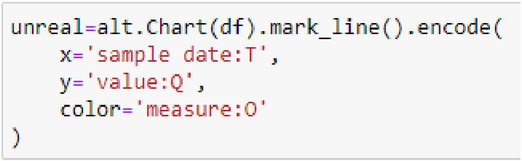


Figure 3



```
iron=unreal.transform_filter(
    (datum.measure=='Iron')
)
manganese=unreal.transform_filter(
    (datum.measure=='Manganese')
)
coliform=unreal.transform_filter(
    (datum.measure=='Total coliforms')
)
alt.hconcat(iron, manganese, coliform)
```

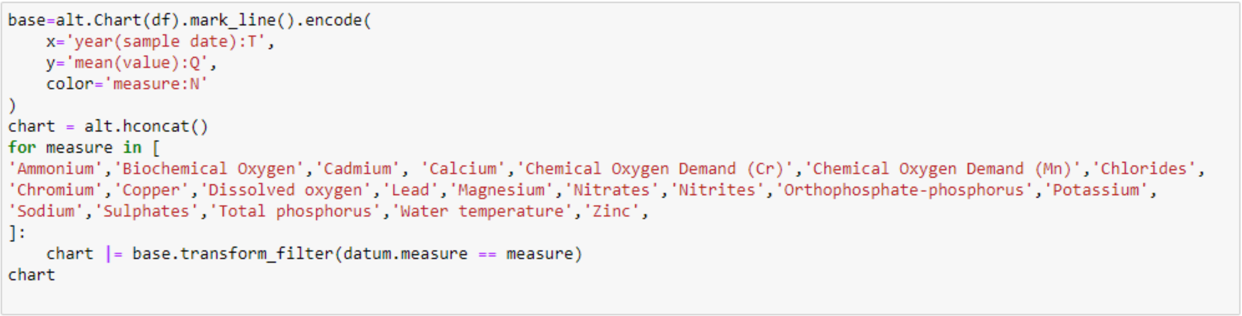


```
unreal=alt.Chart(df).mark_line().encode(
    x='sample date:T',
    y='value:Q',
    color='measure:O'
)
```

For this visualization, I showed the values of certain measures which had unusually high values I found out from the 2nd visualization over time and show how much unusual it is. I used quantitative attributes to length. I used one-dimensional mark : line for it. I think showing how unusual data this visualization was an effective one.

This visualization explicitly shows how unusual the data are.

Visual-4:



```
base=alt.Chart(df).mark_line().encode(
    x='year(sample date):T',
    y='mean(value):Q',
    color='measure:N'
)
chart = alt.hconcat()
for measure in [
'Ammonium','Biochemical Oxygen','Cadmium', 'Calcium','Chemical Oxygen Demand (Cr)','Chemical Oxygen Demand (Mn)','Chlorides',
'Chromium','Copper','Dissolved oxygen','Lead','Magnesium','Nitrates','Nitrites','Orthophosphate-phosphorus','Potassium',
'Sodium','Sulphates','Total phosphorus','Water temperature','Zinc',
]:
    chart |= base.transform_filter(datum.measure == measure)
chart
```

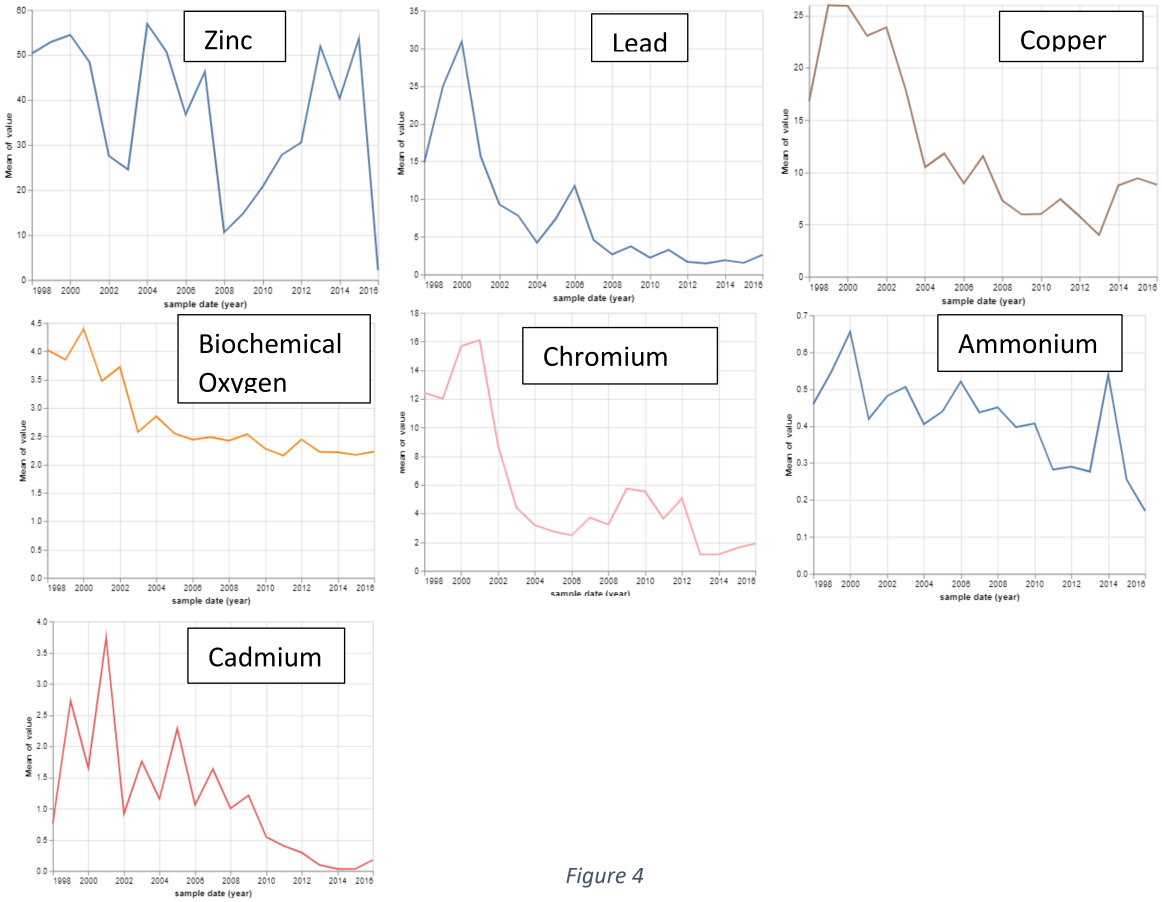
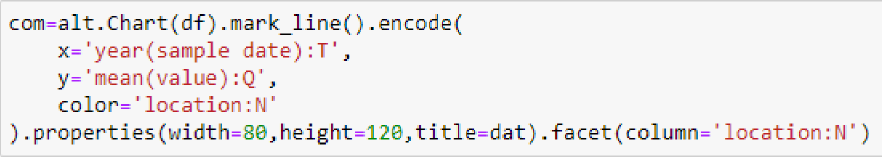


Figure 4

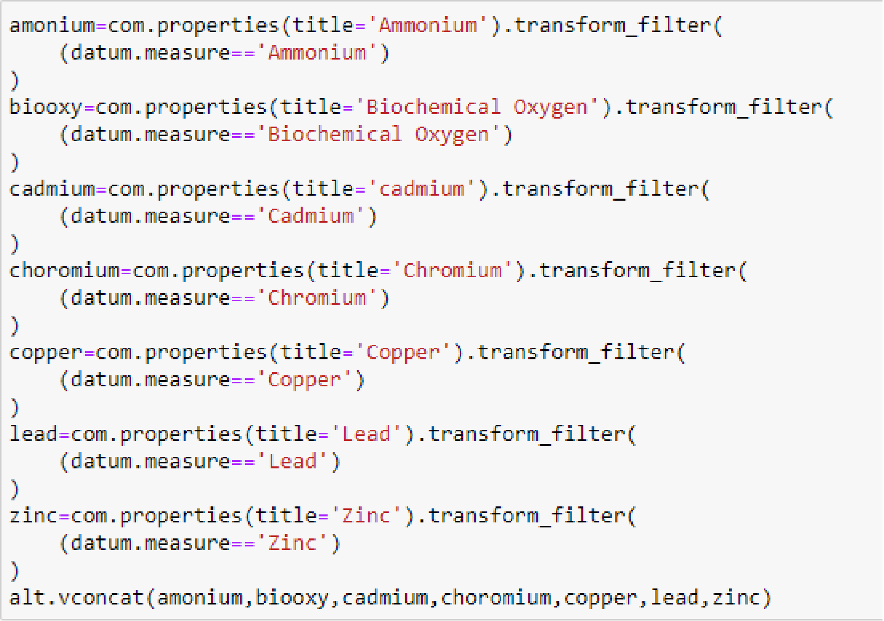
For this visualization, I showed the values of certain measures which I could use to show a pattern over time and see their changes. I used quantitative attributes to length. I used one-dimensional mark : line for it. To identify different marks, I used the color channel. I also transformed the sample data to year and the value to mean value to get a more appropriate pattern. I think to show a pattern of data over time this is a perfect visualization, and I could also find anomalies among them.

I showed a few of them in fig-4. These were the ones with the most change in trend.

Visual-5:



```
com=alt.Chart(df).mark_line().encode(
    x='year(sample date):T',
    y='mean(value):Q',
    color='location:N'
).properties(width=80,height=120,title=dat).facet(column='location:N')
```



```
amonium=com.properties(title='Ammonium').transform_filter(
    (datum.measure=='Ammonium')
)
biooxy=com.properties(title='Biochemical Oxygen').transform_filter(
    (datum.measure=='Biochemical Oxygen')
)
cadmium=com.properties(title='cadmium').transform_filter(
    (datum.measure=='Cadmium')
)
choromium=com.properties(title='Chromium').transform_filter(
    (datum.measure=='Chromium')
)
copper=com.properties(title='Copper').transform_filter(
    (datum.measure=='Copper')
)
lead=com.properties(title='Lead').transform_filter(
    (datum.measure=='Lead')
)
zinc=com.properties(title='Zinc').transform_filter(
    (datum.measure=='Zinc')
)
alt.vconcat(amonium,biooxy,cadmium,choromium,copper,lead,zinc)
```

For this visualization, I picked certain measures which show anomalies in the 4th visualization and showed their values over the years across different locations. I used quantitive attributes to length. I used one-dimensional mark : line for it. To identify different marks, I used the color channel. I also filtered the sample data by year and the value to mean value to get a more appropriate pattern. This chart is good for showing patterns and I also showed it across different locations to identify which location showed it the most.

I showed the entire thing in fig-5. It is easy to compare with this visualization and find out which had the most unusual patterns

Findings

For analysis question 1 my finding was:

i. From my analysis of the dataset, I Found interesting trend for Ammonium, Biochemical Oxygen, Cadmium, Chromium, Lead, Copper, and Zinc. Their values were going down over the years.
ii. I found anomalies in Kohsoom especially as among all the locations Kohsoom had the most chemicals. As shown in the map for the question the waste dumping location was near somewhere Kohsoom. So, it can be assumed that for this reason, the chemical values are so high there.

For the analysis question 2 my finding was:

i. There were some missing values over time for a lot of measures. Which can not be used to do an analysis.
ii. The entire dataset has duplicate entries for the same location, sample date, and measure.
iii. There were some unrealistic values in the dataset. Iron, Total coliforms, and Manganese had high unrealistic values.

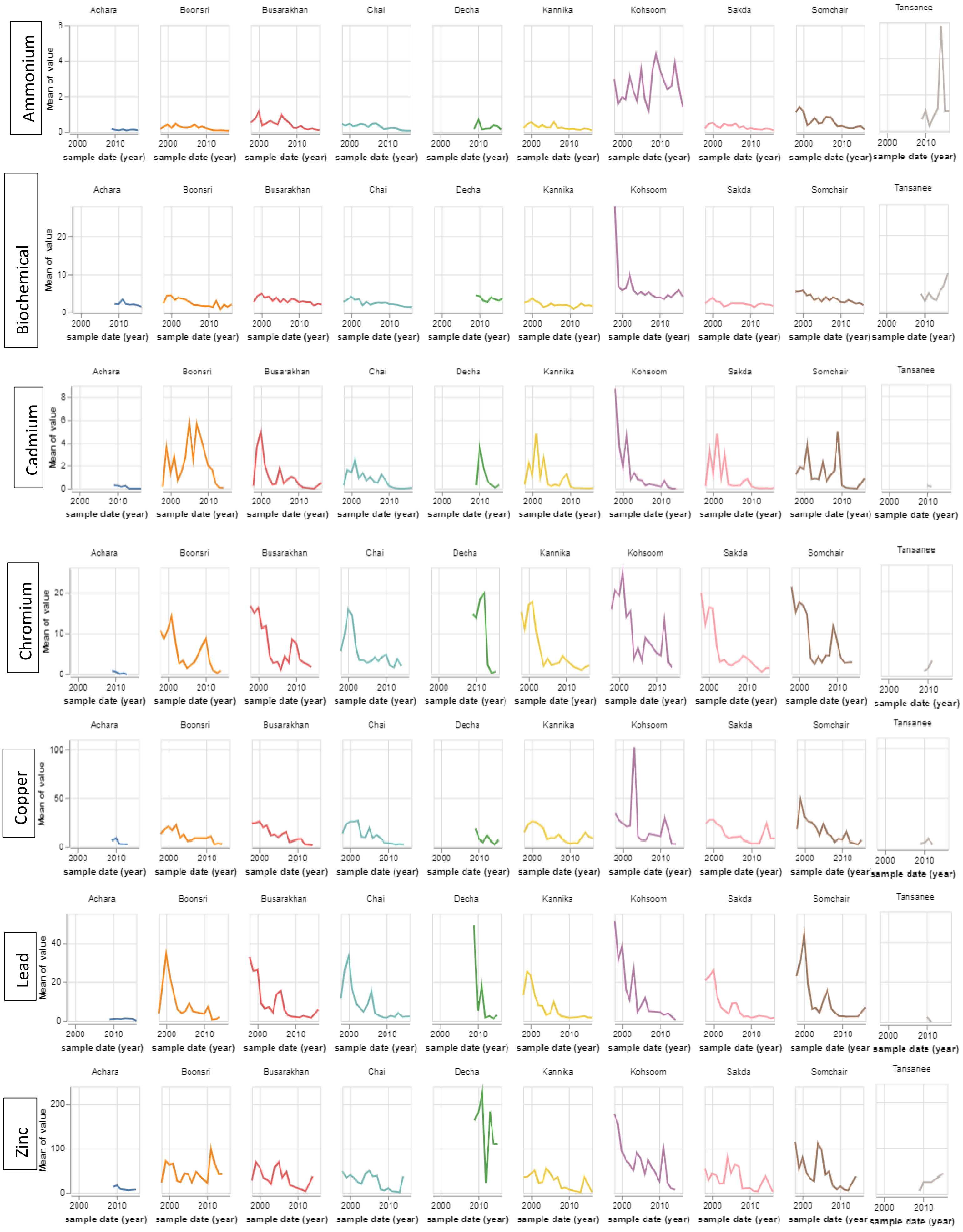

Ammonium
Biochemical
Cadmium
Chromium
Copper
Lead
Zinc
Achara
Boonsri
Busarakhan
Chai
Decha
Kannika
Kohsoom
Sakda
Somchair
Tansanee
Mean of value
sample date (year)
2000
2010
0
2
4
6
10
20
8
50
100
40
200

Figure 5