In [131…
```
import pandas as pd
import altair as alt
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
from IPython.display import Image
from altair.expr import datum
from altair_saver import save

alt.data_transformers.enable('data_server')
pd.set_option('display.max_rows', None)
```

In [132…
```
df = pd.read_csv('Boonsong Lekagul waterways readings.csv',parse_dates=['sample da
df.head()
```

Out[132]:

	id	value	location	sample date	measure
0	2221	2.00	Boonsri	1998-01-11	Water temperature
1	2223	9.10	Boonsri	1998-01-11	Dissolved oxygen
2	2227	0.33	Boonsri	1998-01-11	Ammonium
3	2228	0.01	Boonsri	1998-01-11	Nitrites
4	2229	1.47	Boonsri	1998-01-11	Nitrates

In [133…
```
df['year'] = df['sample date'].dt.year
df['month'] = df['sample date'].dt.month
```

In [ ]:
```
```

In [134…
```
df.head()
```

Out[134]:

	id	value	location	sample date	measure	year	month
0	2221	2.00	Boonsri	1998-01-11	Water temperature	1998	1
1	2223	9.10	Boonsri	1998-01-11	Dissolved oxygen	1998	1
2	2227	0.33	Boonsri	1998-01-11	Ammonium	1998	1
3	2228	0.01	Boonsri	1998-01-11	Nitrites	1998	1
4	2229	1.47	Boonsri	1998-01-11	Nitrates	1998	1

In [135…
```
print('Total unique measurements = ',df['measure'].nunique())
print('Total number of years = ',df['year'].nunique())
print('Total number of months = ',df['month'].nunique())
```

```
Total unique measurements =  106
Total number of years =  19
Total number of months =  12
```

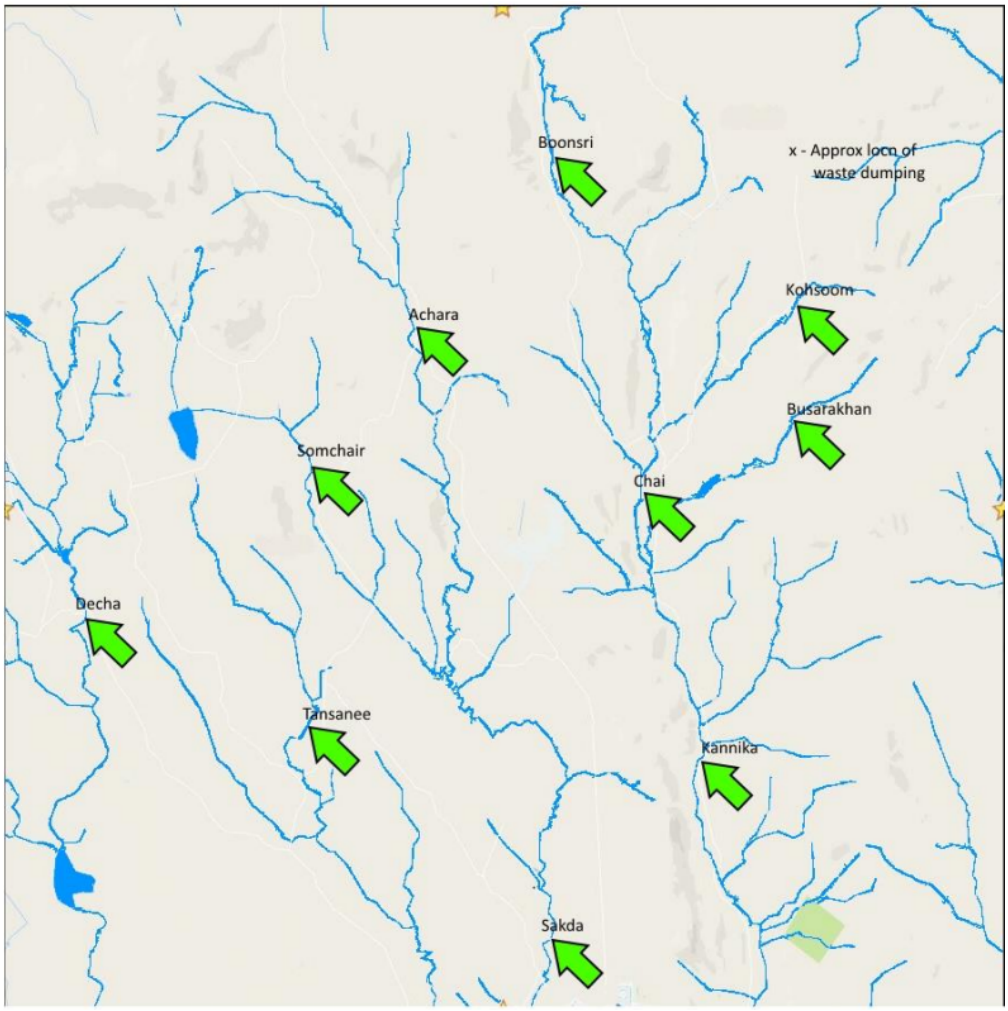
In [136…
```
df.describe().loc[['min', 'max'], ['year', 'month']]
```

Out[136]:

	year	month
min	1998.0	1.0
max	2016.0	12.0

In [137…
```
Image(filename='cst_4060_cw2_map.png', width=500)
```

Out[137]:



Closest locations from dumping site were Bonoosri, Kohsoom and Busarakhan

Q2(i) - Missing Data

In [138…
```
# creating a pivot table in order to get nans for the values that are missing.
missing_data = df.pivot_table(values='value',index='measure',columns='year',aggfun
missing_data.reset_index(inplace=True)
```

In [139…
```
missing_data.head()
```

Out[139]:

year	measure	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
0	1,2,3-Trichlorobenzene	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.001715
1	1,2,4-Trichlorobenzene	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.001000
2	AGOC-3A	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	AOX	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	19.757746
4	Acenaphthene	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.004059

In [140…

```
# Melting it into a format so that we retain the NaN's and we have 3 columns 1 meas
# a Heatmap.
temp_df = pd.melt(missing_data, id_vars=['measure'], value_vars=missing_data.columı
```

In [141…

```
temp_df.head()
```

Out[141]:

	measure	year	value
0	1,2,3-Trichlorobenzene	1998	NaN
1	1,2,4-Trichlorobenzene	1998	NaN
2	AGOC-3A	1998	NaN
3	AOX	1998	0.0
4	Acenaphthene	1998	NaN

In [142…

```
# Looking at all the variables for each year and seeing how many of them are missi

base = alt.Chart(temp_df).mark_rect().encode(x='year:O',
                                             y=alt.Y(field='measure', axis=alt.Axi
                                             tooltip=['year', 'measure', 'value'],
                                             color=alt.Color('value:Q', scale=alt.
    labelFontSize=7.5).properties(height= 800,width=400, title='Missing Values')
base.save('Missing_values.html')
base
```

Out[142]:

In [143…

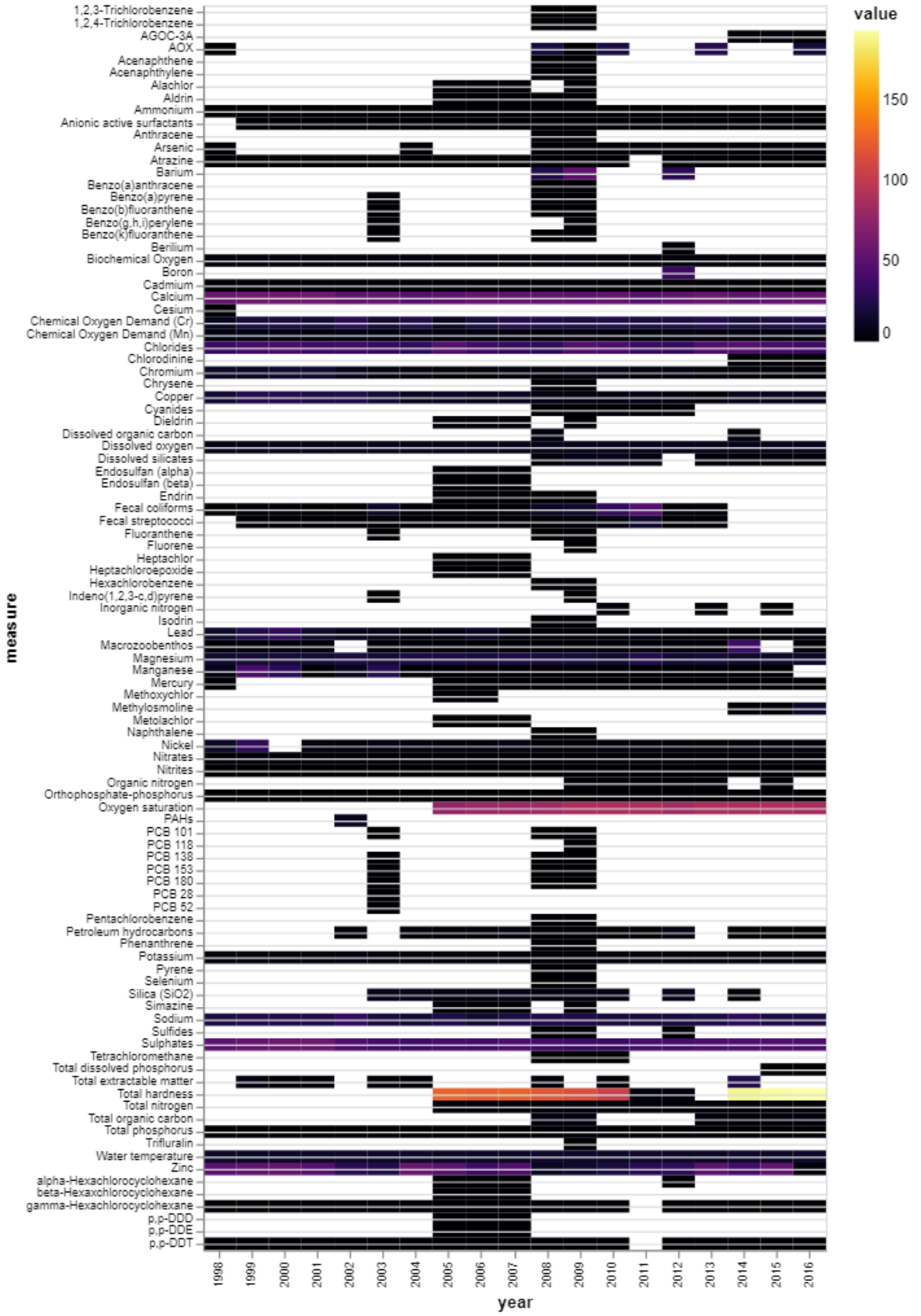
```
temp_df_ur = ['Iron', 'Total dissolved salts', 'Total coliforms', 'Bicarbonates',
for a in temp_df_ur:
    temp_df.drop(temp_df[temp_df['measure'] == a].index, inplace = True)
```

In [144…

```
base = alt.Chart(temp_df).mark_rect().encode(x='year:O',
                                             y=alt.Y(field='measure', axis=alt.Axi
                                             tooltip=['year', 'measure', 'value'],
                                             color=alt.Color('value:Q', scale=alt.
```

```
    labelFontSize=7.5).properties(height= 800,width=400)
base.save('Trend.html')
base
```

Out[144]:



Anomaly

Total Coliforms and aluminium is an anomaly. Macrozoobenthos, Methylosmoline is also and anomaly.

Trends

Total Hardness is a trend, Methylosmoline is a trend for 2 locations.

Q2(ii) - Change in Collection Frequency

In [145…
```
df['measure'].value_counts().sort_values(ascending=False).reset_index().head()
```

Out[145]:

	index	measure
0	Water temperature	5031
1	Nitrites	4791
2	Ammonium	4790
3	Nitrates	4786
4	Orthophosphate-phosphorus	4782

In [146…
```
# All measures with less than 1 year or readings
del_measures = list(df['measure'].value_counts().sort_values(ascending=False).rese
```

In [147…
```
# Deleting all the measures that has less than 365 readings.
for a in del_measures:
    df.drop(df[df['measure'] == a].index, inplace = True)
```

In [148…
```
df.groupby(['year','month'])['value'].count().reset_index().head()
```

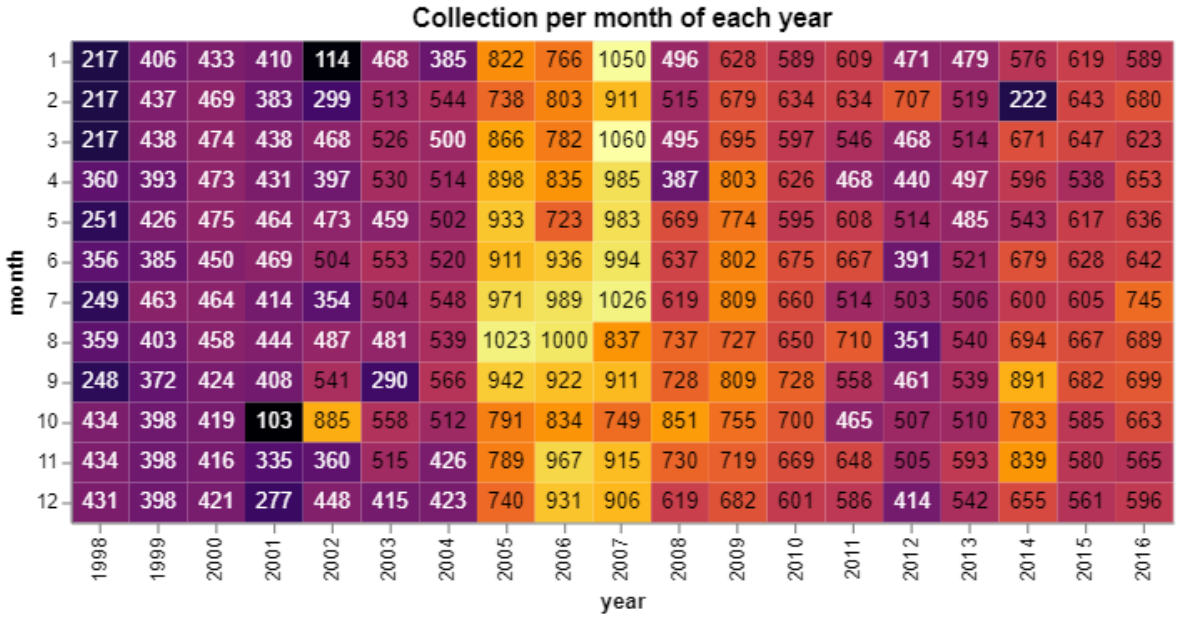
Out[148]:

	year	month	value
0	1998	1	217
1	1998	2	217
2	1998	3	217
3	1998	4	360
4	1998	5	251

In [149…
```
count_freq = df.groupby(['year','month'])['value'].count().reset_index()
```

In [150…
```
temp = alt.Chart(count_freq).mark_rect().encode(
    x='year:O',
    y='month:O',
    color=alt.Color('value:Q', scale=alt.Scale(scheme='inferno'))
).properties(width=550, title='Collection per month of each year')

text = temp.mark_text(baseline='middle').encode(
    text ='value:Q',
    color= alt.condition(datum.value> 500, alt.value('black'), alt.value('white'))
tooltip=['year','month','value'])

(temp+text).save('col_freq.html')
temp+text
```

Out[150]:



In [151…

```
# # Looking at the change in data collection frequency.
# q2_2 = alt.Chart(df).mark_bar().encode(
# x='year',
# y='count(year)',
# tooltip=['year', 'count()']).properties(height=200,width=550, title='Collection

# base = alt.Chart(df).mark_rect().encode(x='year(sample date):O', y='month(sample
#                                         color=alt.Color('count(value):Q', scale=

# #text = base.mark_text(baseline='middle').encode(text ='count(value):Q', color= 
# text = alt.Chart(df).mark_text(baseline='middle').encode(
#     x='year(sample date):O',
#     y='month(sample date):O',
#     text='count()',
#     tooltip=['year','month','count()'],
#     color=alt.condition(datum.text > 100,
#                         alt.value('black'),
#                         alt.value('white'))
# )

# (base+text #& q2_2
# ).save('col_freq.html')
# base+text #& q2_2
```

The peak readings were done in 2007, 2006 and 2005 were almost similar.

In [152…

```
collection_freq = df.pivot_table(values='value',index='measure',columns='year',agg
collection_freq.reset_index(inplace=True)
col_freq = pd.melt(collection_freq, id_vars=['measure'], value_vars=collection_fre
```

In [153…

```
col_freq_diff = df.pivot_table(values='value',index='measure',columns='year',aggfu
col_freq_diff.reset_index(inplace=True)
```

In [154…

```
col_freq_diff = pd.melt(col_freq_diff, id_vars=['measure'], value_vars=col_freq_di
col_freq_diff.head()
```

Out[154]:

	measure	year	value
0	AGOC-3A	1998	NaN
1	AOX	1998	NaN
2	Alachlor	1998	NaN
3	Aldrin	1998	NaN
4	Ammonium	1998	NaN

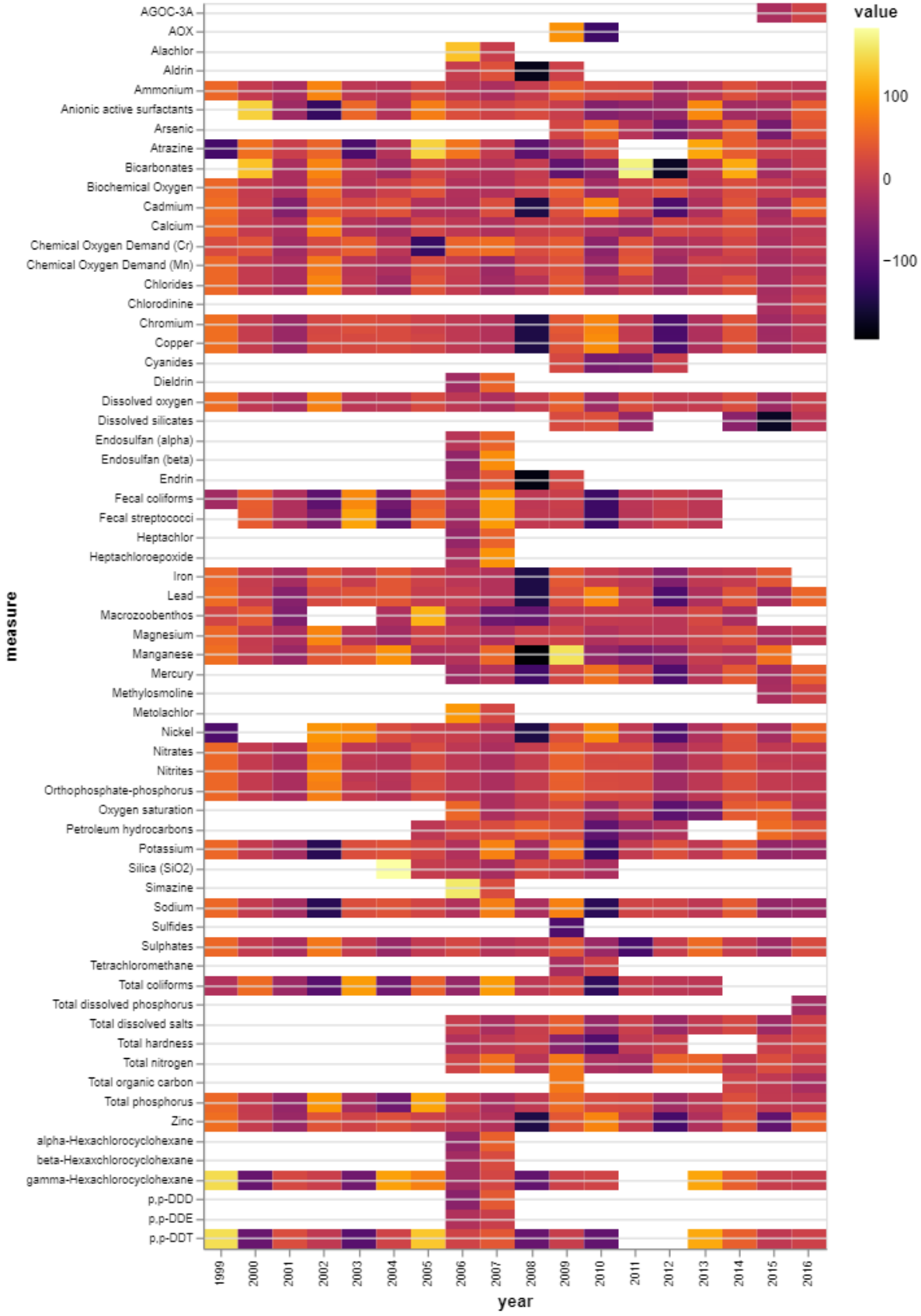
In [155…

```
new_diff = col_freq_diff.loc[col_freq_diff['measure']!='Water temperature', :]
```

In [156…

```
alt.Chart(new_diff).mark_rect().encode(x='year:O',
                                       y=alt.Y(field='measure', axis=alt.Axi
                                       color=alt.Color('value:Q', scale=alt.
    labelFontSize=7.5).properties(height= 800,width=400)
```

Out[156]:



In [157…

```
df[(df['measure']=='Bicarbonates') & (df['year']==2011)].groupby('location').count
```

Out[157]:

location	id	value	sample date	measure	year	month
Boonsri	30	30	30	30	30	30
Busarakhan	12	12	12	12	12	12
Chai	63	63	63	63	63	63
Decha	10	10	10	10	10	10
Kannika	39	39	39	39	39	39
Kohsoom	12	12	12	12	12	12
Sakda	36	36	36	36	36	36
Somchair	12	12	12	12	12	12
Tansanee	12	12	12	12	12	12

In [158…
```
df.shape
```

Out[158]:
```
(133982, 7)
```

In [159…
```
df.head()
```

Out[159]:

	id	value	location	sample date	measure	year	month
0	2221	2.00	Boonsri	1998-01-11	Water temperature	1998	1
1	2223	9.10	Boonsri	1998-01-11	Dissolved oxygen	1998	1
2	2227	0.33	Boonsri	1998-01-11	Ammonium	1998	1
3	2228	0.01	Boonsri	1998-01-11	Nitrites	1998	1
4	2229	1.47	Boonsri	1998-01-11	Nitrates	1998	1

In [160…
```
df['measure'].value_counts().sort_values(ascending=False)
```

Out[160]:

Water temperature	5031
Nitrites	4791
Ammonium	4790
Nitrates	4786
Orthophosphate-phosphorus	4782
Total phosphorus	4600
Dissolved oxygen	4531
Biochemical Oxygen	4488
Manganese	4039
Chlorides	3961
Chemical Oxygen Demand (Mn)	3890
Magnesium	3796
Calcium	3765
Chemical Oxygen Demand (Cr)	3718
Sulphates	3253
Chromium	3015
Lead	3006
Copper	3002
Zinc	2982
Cadmium	2963
Bicarbonates	2826
Total dissolved salts	2789
Iron	2710
gamma-Hexachlorocyclohexane	2580
Total nitrogen	2553
Nickel	2442
Anionic active surfactants	2220
p,p-DDT	2209
Oxygen saturation	2177
Sodium	2171
Potassium	2142
Atrazine	2137
Mercury	1764
Total hardness	1672
Silica (SiO2)	1668
Total coliforms	1615
Fecal coliforms	1490
Fecal streptococci	1301
Macrozoobenthos	1273
Petroleum hydrocarbons	1185
Arsenic	1004
Endrin	710
Aldrin	692
p,p-DDE	661
beta-Hexaxchlorocyclohexane	639
p,p-DDD	632
alpha-Hexachlorocyclohexane	603
Dieldrin	597
Heptachlor	589
Endosulfan (alpha)	566
Dissolved silicates	564
Alachlor	528
Endosulfan (beta)	528
Metolachlor	522
Total dissolved phosphorus	521
AOX	492
Heptachloroepoxide	491
Tetrachloromethane	474
AGOC-3A	474
Methylosmoline	474
Chlorodinine	474
Simazine	441
Total organic carbon	418
Sulfides	389

```
Cyanides                        386
Name: measure, dtype: int64
```

Water temperature, Nitrites, Ammonium are top 3 most frequently readings

In [ ]:

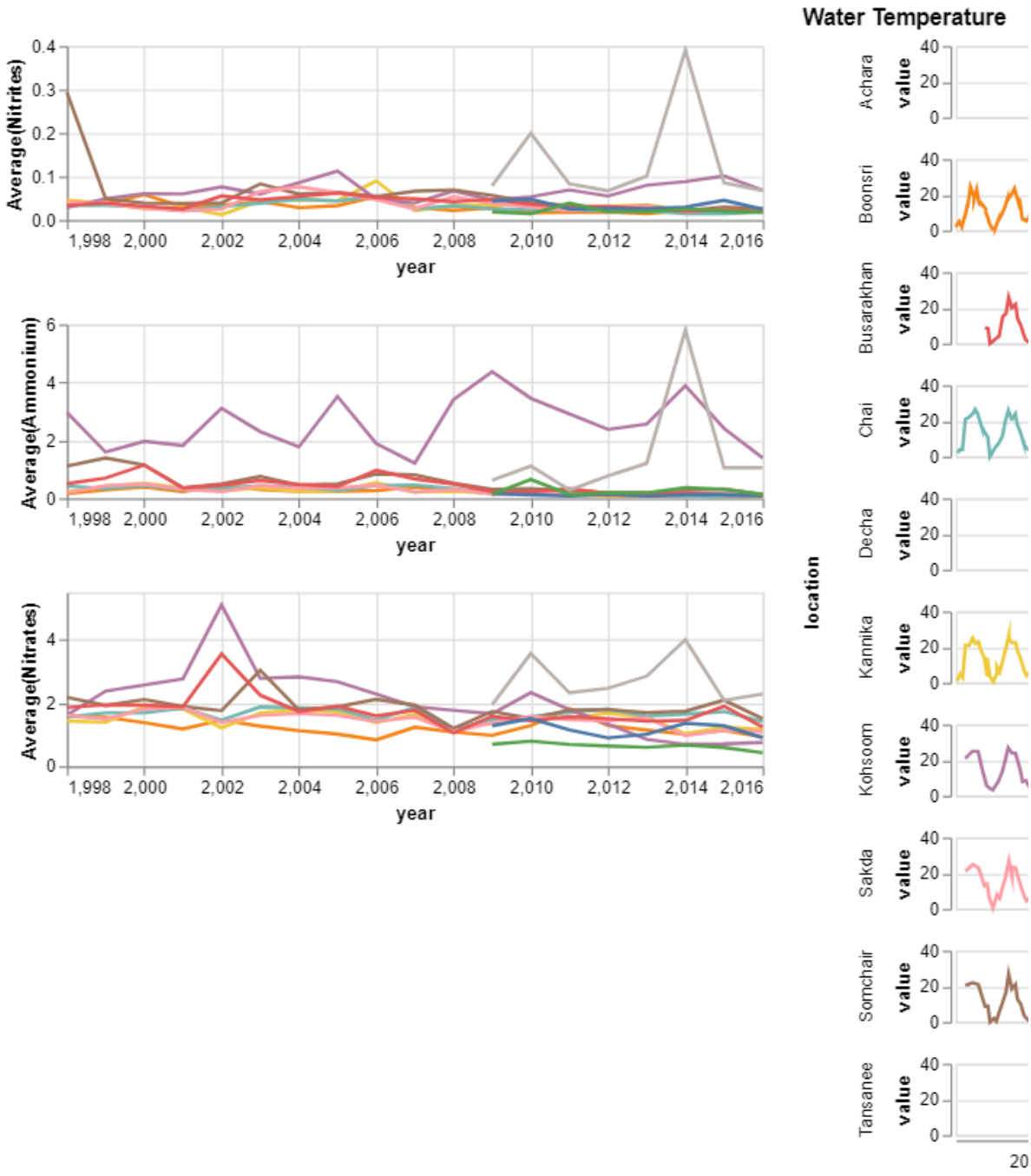
Iron and Macrozoobenthos are unrealistic values

In [161…

```
# Creating a drop down menu for selecting a site that will be used in future plots
input_dropdown = alt.binding_select(options=df['location'].unique())

selection = alt.selection_single(fields=['location'], bind=input_dropdown, name='S

color = alt.condition(selection,alt.Color('location:N'),alt.value('lightgray'))

opacity = alt.condition(selection, alt.value(1.0), alt.value(0.2))
```

In [162…

```
water = alt.Chart(df[df['measure']=='Water temperature']).mark_line().encode(
    x="sample date",
    y="value:Q",
    color="location:N",
    row="location:N"
).properties(
    height=41, width=400, title='Water Temperature'
)

nitrites = alt.Chart(df[df['measure']=='Nitrites']).mark_line().encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average(Nitrites)"),
color=color,
opacity=opacity,
tooltip=['location', 'year', 'mean(value)']).add_selection(selection).properties(h

ammonium = alt.Chart(df[df['measure']=='Ammonium']).mark_line().encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average(Ammonium)"),
color=color,
opacity=opacity,
tooltip=['location', 'year', 'mean(value)']).add_selection(selection).properties(h

nitrates = alt.Chart(df[df['measure']=='Nitrates']).mark_line().encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average(Nitrates)"),
color=color,
opacity=opacity,
tooltip=['location', 'year', 'mean(value)']).add_selection(selection).properties(h

display( (nitrites & ammonium & nitrates) | water)
```



Site_location Boonsri

Tansanee has high readings for Ammonium and Nitrites Kohsoom had high readings for Ammonium

Water readings seem to be very normal according to the seasonal temperature changes

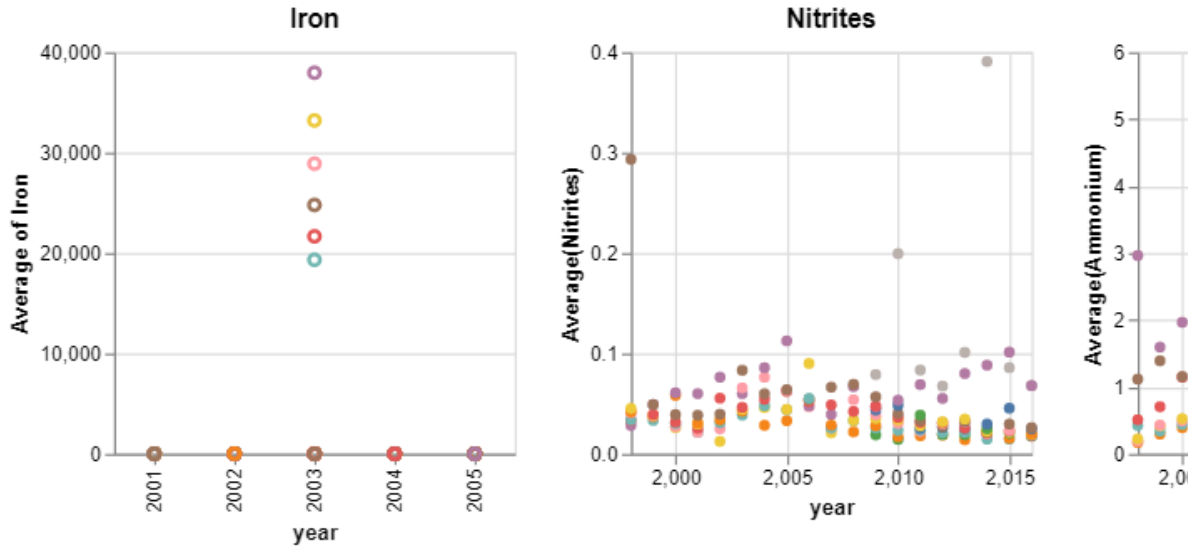
Q2(iii) - Unrealistic Values

In [163…

```
# Plotting unrealistic values.

# Iron that was seen on the first chart of Q2(i).
Iron = alt.Chart(df).mark_point().encode(
alt.X('year:O'),
alt.Y('average(value)', title='Average of Iron'),
color = 'location:N',
order=alt.Order("average(value)", sort="ascending"),
```

```
tooltip=['year', 'location', 'average(value)', 'sample date']).transform_filter(
    (alt.FieldRangePredicate(field='year', range=[2001,2005]))).interactive().prop

Iron_2 = Iron.mark_point().encode().transform_filter(datum.measure=='Iron')

Nitrites = alt.Chart(df).mark_circle().encode(
alt.X(field='year', type='quantitative', scale=alt.Scale(zero=False)),
alt.Y(field='value', aggregate='average', title="Average(Nitrites)"),
color=color,
order=alt.Order("average(value)", sort="ascending"),
tooltip=['year', 'location', 'average(value)']).transform_filter(datum.measure=='N
Nitrites

Ammonium = alt.Chart(df).mark_circle().encode(
alt.X(field='year', type='quantitative', scale=alt.Scale(zero=False)),
alt.Y(field='value', aggregate='average', title="Average(Ammonium)"),
color=color,
order=alt.Order("average(value)", sort="ascending"),
tooltip=['year', 'location', 'average(value)']).transform_filter(datum.measure=='A
Ammonium


(Iron_2 | Nitrites | Ammonium).save('unrealistic_values.html')
Iron_2 | Nitrites | Ammonium
```

Out[163]:



Site_location Boonsri

Q1 - Describe trends and anomalies with respect to chemical contamination

First we will create a pivot table and calculate the difference of each measure for each year. And we will note down the ones with interesting change in values.

In [164...

```
df.pivot_table(values='value',index='measure',columns='year',aggfunc=np.mean).diff
```

Out[164]:

year	1998	1999	2000	2001	2002	2003	
measure							
AGOC-3A	NaN	NaN	NaN	NaN	NaN	NaN	
AOX	NaN	NaN	NaN	NaN	NaN	NaN	
Alachlor	NaN	NaN	NaN	NaN	NaN	NaN	
Aldrin	NaN	NaN	NaN	NaN	NaN	NaN	
Ammonium	NaN	0.088754	0.106499	-0.236374	0.062330	0.024617	-0.1
Anionic active surfactants	NaN	NaN	0.020424	0.015743	-0.022000	-0.040264	-0.0
Arsenic	NaN	NaN	NaN	NaN	NaN	NaN	
Atrazine	NaN	0.043333	-0.001542	-0.021220	-0.020571	0.000000	0.0
Bicarbonates	NaN	NaN	-28.317349	4.000122	-21.362773	27.852365	-4.2
Biochemical Oxygen	NaN	-0.170307	0.539044	-0.918661	0.248298	-1.148961	0.2
Cadmium	NaN	1.963756	-1.072325	2.077416	-2.814854	0.844673	-0.6
Calcium	NaN	4.091652	-0.810030	0.592862	-7.586775	0.873265	-5.9
Chemical Oxygen Demand (Cr)	NaN	10.940265	0.924842	-3.753684	9.851578	-10.660118	7.7
Chemical Oxygen Demand (Mn)	NaN	1.626648	-0.257130	-1.437695	0.998637	-0.582876	-0.4
Chlorides	NaN	5.645477	2.325618	-5.264993	1.202648	-4.628547	-1.6
Chlorodinine	NaN	NaN	NaN	NaN	NaN	NaN	
Chromium	NaN	-0.409177	3.671343	0.421843	-7.349438	-4.339947	-1.2
Copper	NaN	9.191330	-0.066923	-2.846195	0.792938	-5.959127	-7.4
Cyanides	NaN	NaN	NaN	NaN	NaN	NaN	
Dieldrin	NaN	NaN	NaN	NaN	NaN	NaN	
Dissolved oxygen	NaN	0.569807	0.364576	-0.158142	0.074597	0.318632	-0.3
Dissolved silicates	NaN	NaN	NaN	NaN	NaN	NaN	
Endosulfan (alpha)	NaN	NaN	NaN	NaN	NaN	NaN	
Endosulfan (beta)	NaN	NaN	NaN	NaN	NaN	NaN	
Endrin	NaN	NaN	NaN	NaN	NaN	NaN	
Fecal coliforms	NaN	0.948465	0.530762	-1.116767	-0.183891	12.081618	-11.4
Fecal streptococci	NaN	NaN	-0.318362	-0.048757	-0.006383	0.828811	-0.7
Heptachlor	NaN	NaN	NaN	NaN	NaN	NaN	
Heptachloroepoxide	NaN	NaN	NaN	NaN	NaN	NaN	
Iron	NaN	0.120833	0.056322	0.015083	0.312084	1102.183056	-1102.8
Lead	NaN	10.179204	5.849893	-15.193284	-6.400879	-1.493709	-3.5
Macrozoobenthos	NaN	1.266341	-0.014204	0.003483	NaN	NaN	-0.2
Magnesium	NaN	1.754098	1.470401	-1.443208	3.737417	-3.308630	3.3

year	1998	1999	2000	2001	2002	2003	
measure							
Manganese	NaN	33.325779	-12.791827	-28.498298	8.645219	17.532246	-26.0
Mercury	NaN	NaN	NaN	NaN	NaN	NaN	
Methylosmoline	NaN	NaN	NaN	NaN	NaN	NaN	
Metolachlor	NaN	NaN	NaN	NaN	NaN	NaN	
Nickel	NaN	17.791667	NaN	NaN	-2.468170	3.030778	-1.0
Nitrates	NaN	0.071792	0.155129	-0.013971	-0.068002	0.057541	-0.1
Nitrites	NaN	-0.013504	-0.001540	-0.008026	0.002675	0.016575	0.0
Orthophosphate-phosphorus	NaN	-0.014121	0.002245	-0.008106	0.003736	-0.005562	-0.0
Oxygen saturation	NaN	NaN	NaN	NaN	NaN	NaN	
Petroleum hydrocarbons	NaN	NaN	NaN	NaN	NaN	NaN	
Potassium	NaN	0.551920	-0.060785	-0.350889	1.830734	-1.778755	-0.6
Silica (SiO2)	NaN	NaN	NaN	NaN	NaN	NaN	-0.0
Simazine	NaN	NaN	NaN	NaN	NaN	NaN	
Sodium	NaN	2.902061	-0.113537	-0.266608	8.327803	-12.046337	-2.8
Sulfides	NaN	NaN	NaN	NaN	NaN	NaN	
Sulphates	NaN	4.393909	2.737904	0.500439	-13.901445	-5.628554	-2.0
Tetrachloromethane	NaN	NaN	NaN	NaN	NaN	NaN	
Total coliforms	NaN	3.079208	9.086673	-9.916328	-0.193706	46.990609	-42.2
Total dissolved phosphorus	NaN	NaN	NaN	NaN	NaN	NaN	
Total dissolved salts	NaN	NaN	NaN	NaN	NaN	NaN	
Total hardness	NaN	NaN	NaN	NaN	NaN	NaN	
Total nitrogen	NaN	NaN	NaN	NaN	NaN	NaN	
Total organic carbon	NaN	NaN	NaN	NaN	NaN	NaN	
Total phosphorus	NaN	-0.033462	0.005809	-0.013682	0.000612	0.002891	0.0
Water temperature	NaN	-0.889303	0.029312	1.058139	0.455889	-0.646079	0.9
Zinc	NaN	2.587313	1.526165	-6.057564	-20.779085	-3.078780	32.4
alpha-Hexachlorocyclohexane	NaN	NaN	NaN	NaN	NaN	NaN	
beta-Hexaxchlorocyclohexane	NaN	NaN	NaN	NaN	NaN	NaN	
gamma-Hexachlorocyclohexane	NaN	0.043893	-0.004796	0.087146	-0.109181	-0.006144	0.0
p,p-DDD	NaN	NaN	NaN	NaN	NaN	NaN	
p,p-DDE	NaN	NaN	NaN	NaN	NaN	NaN	
p,p-DDT	NaN	0.149512	0.077547	0.140640	0.199831	0.006654	0.0

Following measures seem interesting via the above table.

Arsenic, Lead, Total hardness, Zinc, Chlorides, Fecal coliforms, Manganese, Methylosmoline, Total coliforms, Total dissolved salts

Let's look at them even further

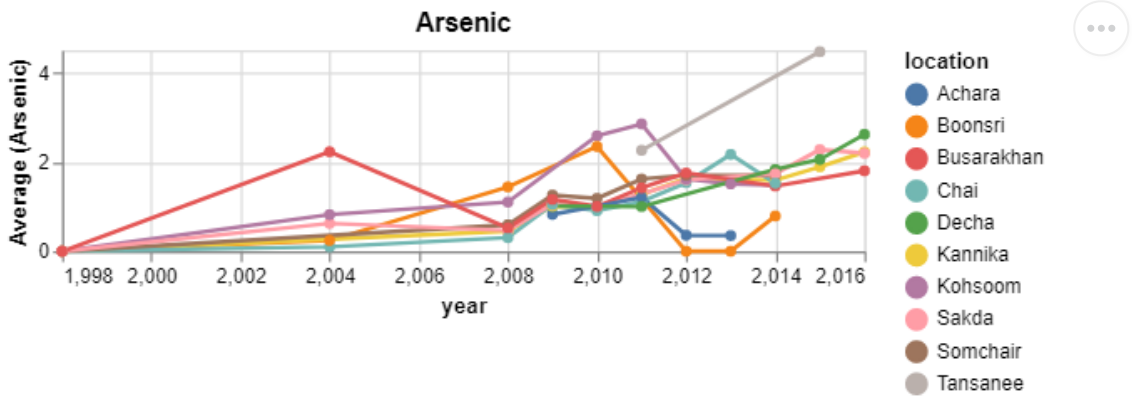
In [165…
```
global_color = alt.Color('location:N')
```

In [166…
```
# # Creating a drop down menu for selecting a site that will be used in future plot
# input_dropdown = alt.binding_select(options=df['location'].unique())

# selection = alt.selection_single(fields=['location'], bind=input_dropdown, name=

# color = alt.condition(selection,alt.Color('location:N'),alt.value('lightgray'))

# opacity = alt.condition(selection, alt.value(1.0), alt.value(0.5))
```

Q1(i) - Trends

In [167…
```
Arsenic = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Arsenic) "),
color=color,
opacity=opacity,
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Arsenic').add_selection(selection).properties(height=100, widtl
Arsenic
```

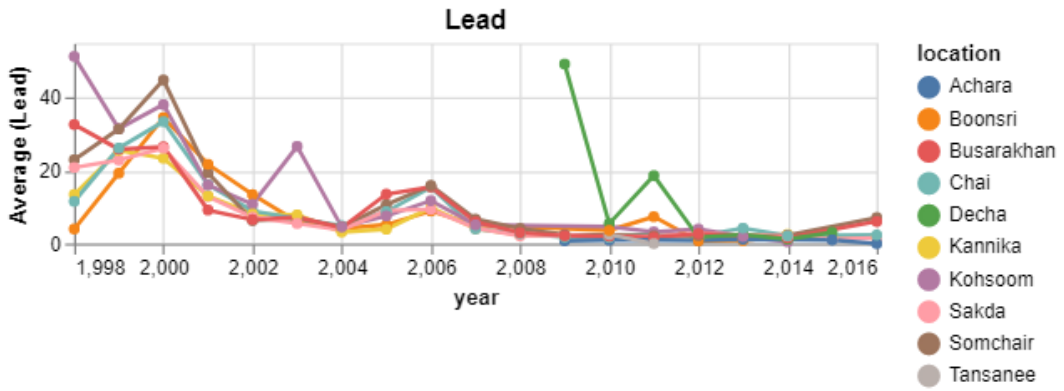
Out[167]:



Site_location Boonsri

In [168…
```
Lead = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Lead) "),
color=color,
opacity=opacity,
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Lead').add_selection(selection).properties(height=100, width=4
Lead
```

Out[168]:

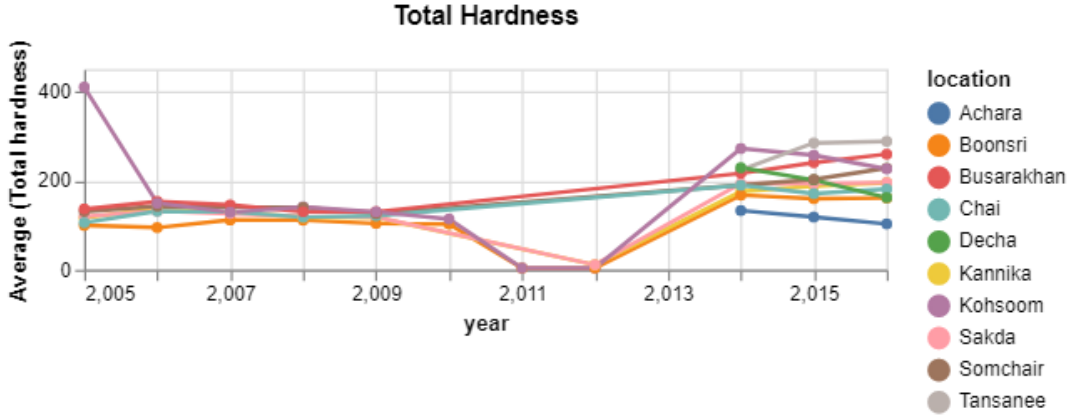


Site_location Boonsri

In [169…

```
Total_hardness = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Total hardness) "),
color=color,
opacity=opacity,
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Total hardness').add_selection(selection).properties(height=10
Total_hardness
```

Out[169]:

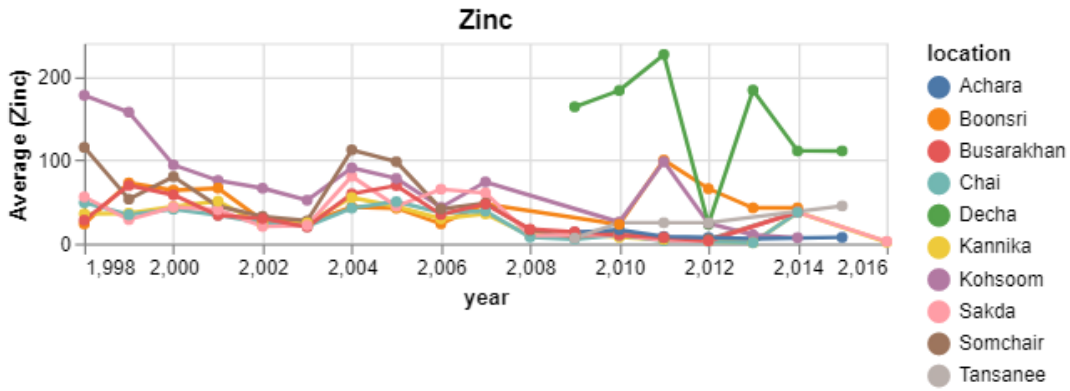


Site_location Boonsri

In [170…

```
Zinc = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Zinc) "),
color=color,
opacity=opacity,
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Zinc').add_selection(selection).properties(height=100, width=4

Zinc
```

Out[170]:



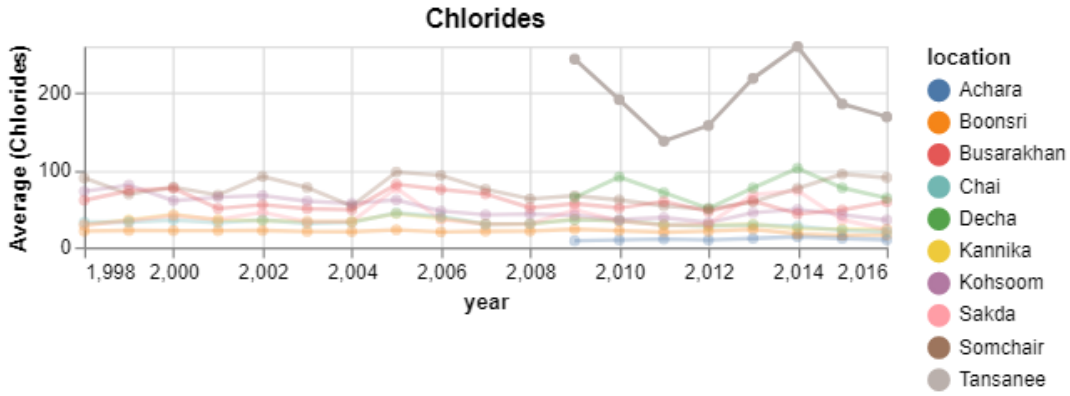
Site_location Boonsri

Q1(ii) - Anomalies

In [171…

```
Chlorides = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Chlorides) "),
color='location:N',
opacity=alt.condition(datum.location=='Tansanee', alt.value(1), alt.value(0.3)),
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Chlorides').properties(height=100, width=400, title='Chlorides
Chlorides
```

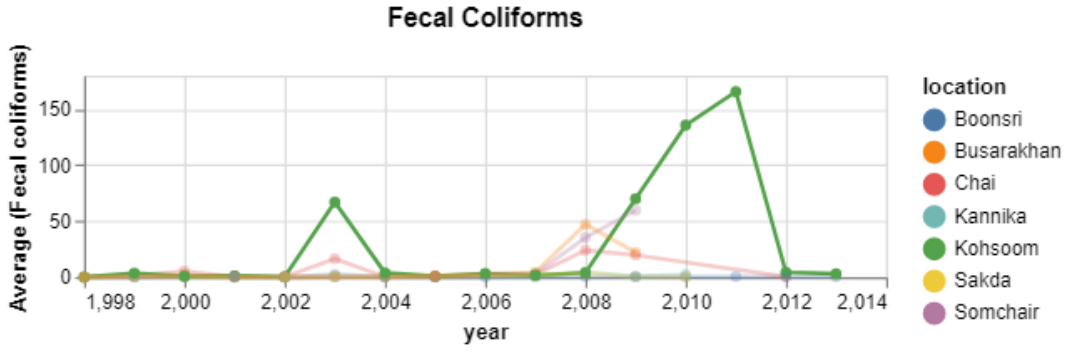
Out[171]:



In [172…

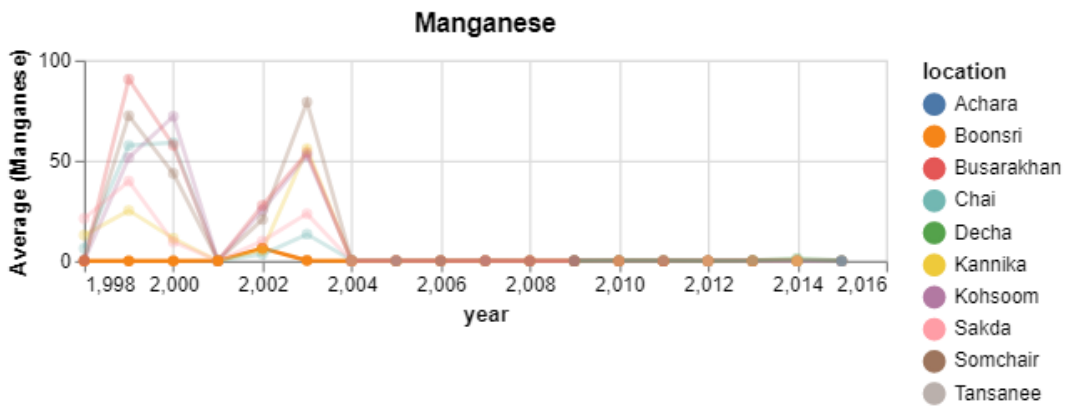
```
Fecal_coliforms = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Fecal coliforms) "),
color='location:N',
opacity=alt.condition(datum.location=='Kohsoom', alt.value(1), alt.value(0.3)),
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Fecal coliforms').properties(height=100, width=400, title='Feca
Fecal_coliforms
```

Out[172]:



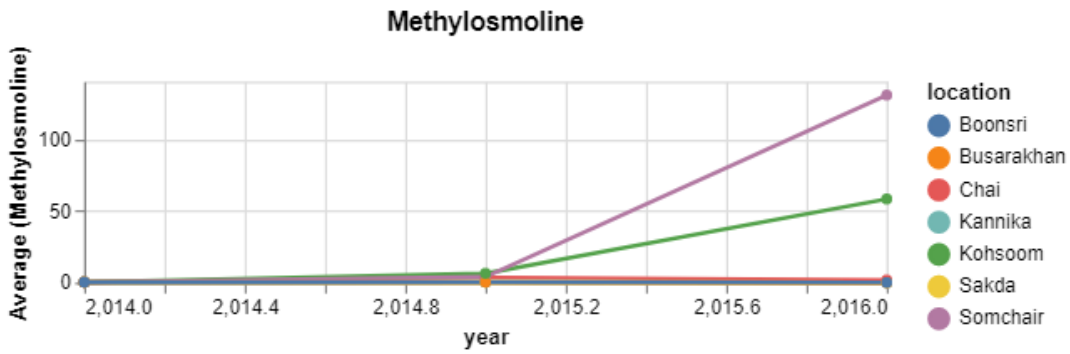
In [173…
```
Manganese = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Manganese) "),
color='location:N',
opacity=alt.condition(datum.location=='Boonsri', alt.value(1), alt.value(0.3)),
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Manganese').properties(height=100, width=400, title='Manganese
Manganese
```

Out[173]:



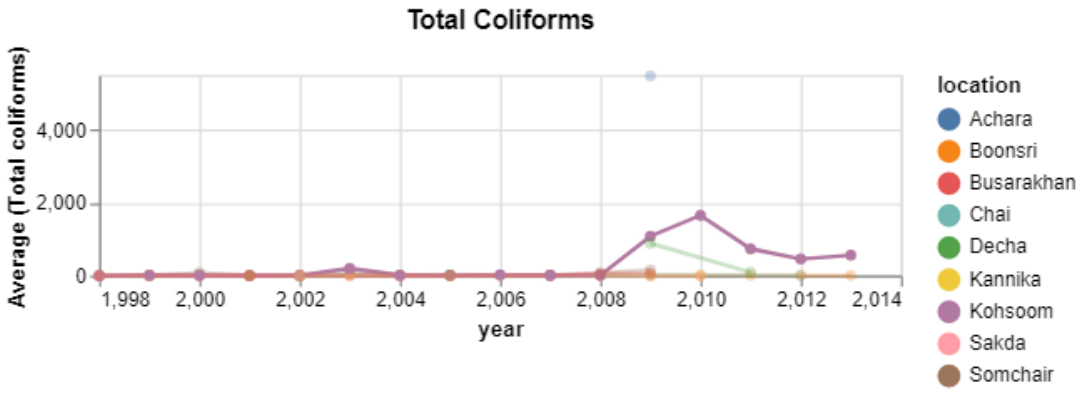
In [174…
```
Methylosmoline = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Methylosmoline) "),
color='location:N',
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Methylosmoline').properties(height=100, width=400, title='Methy
Methylosmoline
```

Out[174]:



In [175…
```
Total_coliforms = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Total coliforms) "),
color='location:N',
opacity=alt.condition(datum.location=='Kohsoom', alt.value(1), alt.value(0.3)),
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Total coliforms').properties(height=100, width=400, title='Tota
Total_coliforms
```

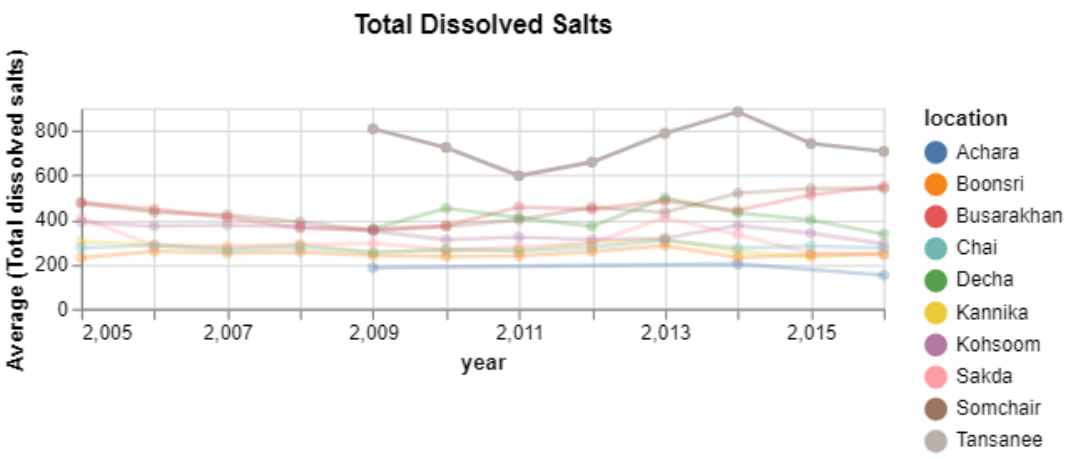
Out[175]:



In [176…

```
Total_dissolved_salts = alt.Chart(df).mark_line(point=True).encode(
alt.X(field='year', type='quantitative'),
alt.Y(field='value', aggregate='average', title="Average (Total dissolved salts) "
color='location:N',
opacity=alt.condition(datum.location=='Tansanee', alt.value(1), alt.value(0.3)),
tooltip=['location', 'year', 'mean(value)']).transform_filter(
    datum.measure=='Total dissolved salts').properties(height=100, width=400, titl
Total_dissolved_salts
```

Out[176]:



Trend - Arsenic (increasing trend), Lead (Decreasing Trend), Total hardness (Increasing Trend), Zinc(Decreasing Trend)

Anomaly - Chlorides (Tanasee different), Fecal coliforms (Kohsoom Different), Manganese(M Shaped starting trend for all but Boonsri), Methylosmoline (Kohsoom and Somchair different), Total coliforms (Kohsoom different), Total dissolved salts (Tanasee different)

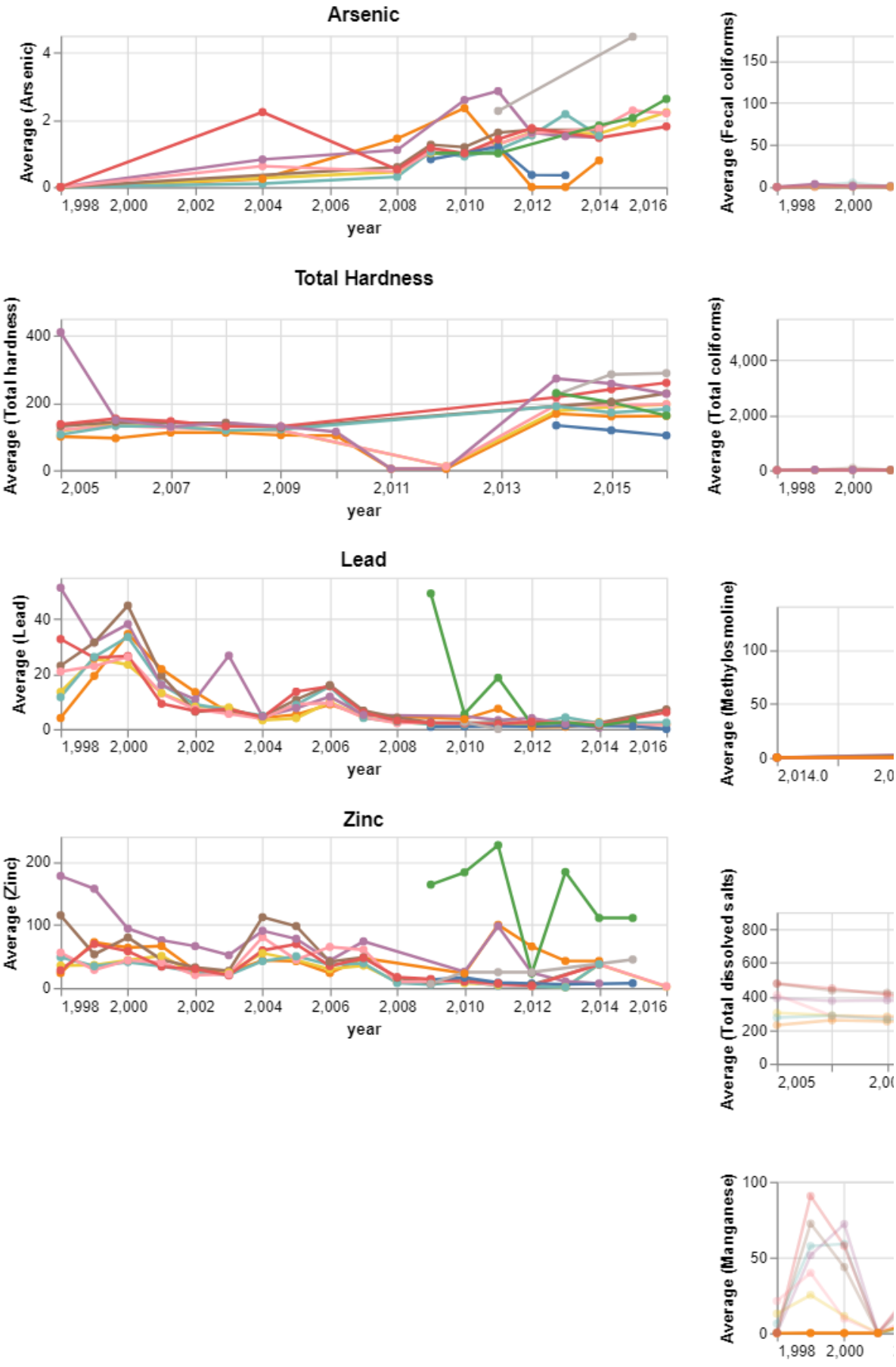
In [177…

```
((Arsenic & Total_hardness & Lead & Zinc) | (Fecal_coliforms & Total_coliforms & M
```

In [178…

```
print('\033[1m' + '   Trends                                                    A
display( (Arsenic & Total_hardness & Lead & Zinc) | (Fecal_coliforms & Total_colif
```

Trends **Anomalies**

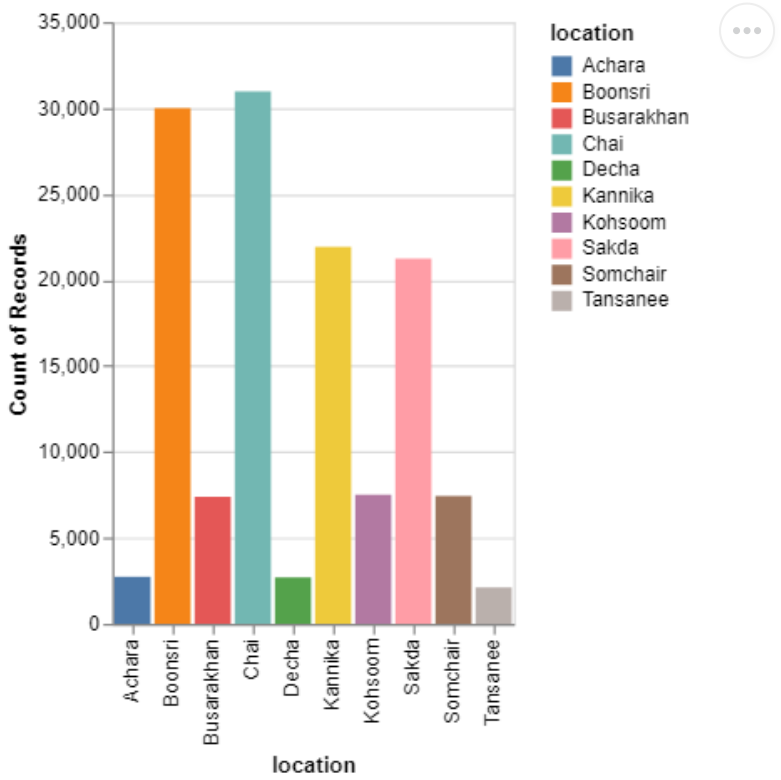


Trying to find more information

In [179…
```
# Number of obs per location
```

In [180…
```
alt.Chart(df).mark_bar().encode(
alt.X("location:N"),
y='count()',
tooltip='count():Q',
color='location:N')
```

Out[180]:



In [181…
```
# Another attempt to find more measures. By Calculating the % of change from 2015 t
```

In [182…
```
df_heatmap = df.pivot_table(values='value',index='measure',columns='year',aggfunc=
```

In [183…
```
df_heatmap['% diff'] = ( df_heatmap[2016] - df_heatmap[2015] ) / df_heatmap[2015]
```

In [184…
```
df_heatmap
```

Out[184]:

year	1998	1999	2000	2001	2002	2003
measure						
AGOC-3A	NaN	NaN	NaN	NaN	NaN	NaN
AOX	0.000000	NaN	NaN	NaN	NaN	NaN
Alachlor	NaN	NaN	NaN	NaN	NaN	NaN
Aldrin	NaN	NaN	NaN	NaN	NaN	NaN
Ammonium	0.460167	0.548920	0.655419	0.419045	0.481375	0.505992
Anionic active surfactants	NaN	0.065833	0.086257	0.102000	0.080000	0.039736
Arsenic	0.000000	NaN	NaN	NaN	NaN	NaN
Atrazine	0.000000	0.043333	0.041791	0.020571	0.000000	0.000000
Bicarbonates	NaN	197.719583	169.402235	173.402357	152.039583	179.891948
Biochemical Oxygen	4.024000	3.853693	4.392737	3.474076	3.722374	2.573413
Cadmium	0.760244	2.724000	1.651675	3.729091	0.914237	1.758910
Calcium	56.940167	61.031818	60.221788	60.814650	53.227875	54.101140
Chemical Oxygen Demand (Cr)	9.041597	19.981862	20.906704	17.153020	27.004598	16.344480
Chemical Oxygen Demand (Mn)	5.018750	6.645398	6.388268	4.950573	5.949211	5.366335
Chlorides	34.406000	40.051477	42.377095	37.112102	38.314750	33.686203
Chlorodinine	NaN	NaN	NaN	NaN	NaN	NaN
Chromium	12.417886	12.008710	15.680052	16.101895	8.752457	4.412510
Copper	16.795122	25.986452	25.919529	23.073333	23.866271	17.907144
Cyanides	NaN	NaN	NaN	NaN	NaN	NaN
Dieldrin	NaN	NaN	NaN	NaN	NaN	NaN
Dissolved oxygen	7.663950	8.233757	8.598333	8.440191	8.514788	8.833420
Dissolved silicates	NaN	NaN	NaN	NaN	NaN	NaN
Endosulfan (alpha)	NaN	NaN	NaN	NaN	NaN	NaN
Endosulfan (beta)	NaN	NaN	NaN	NaN	NaN	NaN
Endrin	NaN	NaN	NaN	NaN	NaN	NaN
Fecal coliforms	0.003357	0.951822	1.482584	0.365817	0.181926	12.263544
Fecal streptococci	NaN	0.380677	0.062314	0.013557	0.007174	0.835985
Heptachlor	NaN	NaN	NaN	NaN	NaN	NaN
Heptachloroepoxide	NaN	NaN	NaN	NaN	NaN	NaN
Iron	0.890833	1.011667	1.067989	1.083072	1.395156	1103.578212
Lead	14.849593	25.028798	30.878691	15.685407	9.284528	7.790820
Macrozoobenthos	0.792500	2.058841	2.044637	2.048120	NaN	2.229891
Magnesium	15.945333	17.699432	19.169832	17.726624	21.464042	18.155411

year	1998	1999	2000	2001	2002	2003
measure						
Manganese	8.058398	41.384177	28.592351	0.094052	8.739271	26.271517
Mercury	0.000000	NaN	NaN	NaN	NaN	NaN
Methylosmoline	NaN	NaN	NaN	NaN	NaN	NaN
Metolachlor	NaN	NaN	NaN	NaN	NaN	NaN
Nickel	12.608333	30.400000	NaN	4.911000	2.442830	5.473608
Nitrates	1.568833	1.640625	1.795754	1.781783	1.713782	1.771322
Nitrites	0.052083	0.038580	0.037039	0.029013	0.031688	0.048263
Orthophosphate-phosphorus	0.074167	0.060045	0.062291	0.054185	0.057921	0.052359
Oxygen saturation	NaN	NaN	NaN	NaN	NaN	NaN
Petroleum hydrocarbons	NaN	NaN	NaN	NaN	1.852857	NaN
Potassium	3.698250	4.250170	4.189385	3.838497	5.669231	3.890476
Silica (SiO2)	NaN	NaN	NaN	NaN	NaN	9.236458
Simazine	NaN	NaN	NaN	NaN	NaN	NaN
Sodium	21.365667	24.267727	24.154190	23.887582	32.215385	20.169048
Sulfides	NaN	NaN	NaN	NaN	NaN	NaN
Sulphates	51.104500	55.498409	58.236313	58.736752	44.835307	39.206753
Tetrachloromethane	NaN	NaN	NaN	NaN	NaN	NaN
Total coliforms	0.122704	3.201912	12.288585	2.372258	2.178552	49.169160
Total dissolved phosphorus	NaN	NaN	NaN	NaN	NaN	NaN
Total dissolved salts	NaN	NaN	NaN	NaN	NaN	NaN
Total hardness	NaN	NaN	NaN	NaN	NaN	NaN
Total nitrogen	NaN	NaN	NaN	NaN	NaN	NaN
Total organic carbon	NaN	NaN	NaN	NaN	NaN	NaN
Total phosphorus	0.130167	0.096705	0.102514	0.088832	0.089444	0.092336
Water temperature	13.765546	12.876243	12.905556	13.963694	14.419583	13.773504
Zinc	50.323171	52.910484	54.436649	48.379085	27.600000	24.521220
alpha-Hexachlorocyclohexane	NaN	NaN	NaN	NaN	NaN	NaN
beta-Hexaxchlorocyclohexane	NaN	NaN	NaN	NaN	NaN	NaN
gamma-Hexachlorocyclohexane	0.002364	0.046257	0.041461	0.128607	0.019426	0.013283
p,p-DDD	NaN	NaN	NaN	NaN	NaN	NaN
p,p-DDE	NaN	NaN	NaN	NaN	NaN	NaN
p,p-DDT	0.003500	0.153012	0.075465	0.216105	0.016274	0.022929

In [185…
```
# Keeping only the measures who have more than 99% increase or decrease
```

In [186…
```
df_heatmap = df_heatmap[(df_heatmap['% diff']>99) | (df_heatmap['% diff'] < -99)]
df_heatmap
```

Out[186]:

year	1998	1999	2000	2001	2002	2003	2004	2005	20
measure									
Cadmium	0.760244	2.724	1.651675	3.729091	0.914237	1.75891	1.154496	2.286954	1.0585
Methylosmoline	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
Petroleum hydrocarbons	NaN	NaN	NaN	NaN	1.852857	NaN	1.525857	0.765258	1.5374

Methlyosmoline can only be considered here because the other 2 measures value in '2016' is either similar or lower then it's values from last decade.

Methylosmoline has a sharp increase and we already have used it.

In [187…
```
corr = {}
```

In [188…
```
#Trying to find measures that are highly Correlated (over 50)
for a in df['measure'].unique():
    temp_df = df[(df['measure']==a) & (df['year']>2009)]
    temp_df.groupby('year')['value'].sum().reset_index()
    corr[a] = temp_df['year'].corr(temp_df['value'])
```

In [189…
```
corr
```

Out[189]:

```
{'Water temperature': 0.009460191590911325,
 'Dissolved oxygen': -0.005276411494355048,
 'Ammonium': -0.04120332794932087,
 'Nitrites': -0.06676094443044986,
 'Nitrates': -0.17430269609961943,
 'Orthophosphate-phosphorus': -0.0854581791557303,
 'Total phosphorus': -0.124299406788089,
 'Sodium': 0.08187389410239546,
 'Potassium': 0.0441600268847656,
 'Calcium': 0.05184766430207902,
 'Magnesium': -0.2319569876470511,
 'Chlorides': 0.013777656341858239,
 'Sulphates': 0.14462298253984732,
 'Iron': 0.03990717395247964,
 'Manganese': 0.06191835626134894,
 'Zinc': 0.015710606270894777,
 'Copper': 0.10636010083131593,
 'Chromium': -0.29686729134761375,
 'Lead': -0.05950267075754541,
 'Cadmium': -0.20118173114084445,
 'Mercury': -0.34323096995015956,
 'Nickel': -0.019143946302640776,
 'Arsenic': 0.2917183818972909,
 'Biochemical Oxygen': -0.012337936646584351,
 'Chemical Oxygen Demand (Cr)': 0.01039819856463897,
 'Chemical Oxygen Demand (Mn)': 0.0836102549897286,
 'AOX': -0.16989048328903236,
 'Atrazine': -0.8036555215998982,
 'Macrozoobenthos': 0.288552509542875,
 'Total coliforms': -0.09433555428133873,
 'Fecal coliforms': -0.14540965124938765,
 'p,p-DDT': -0.07993257219223025,
 'gamma-Hexachlorocyclohexane': -0.7528786840318084,
 'Bicarbonates': -0.13255407980125128,
 'Anionic active surfactants': 0.1150088211704779,
 'Fecal streptococci ': -0.09475657586148993,
 'Petroleum hydrocarbons': -0.14357865447096613,
 'Silica (SiO2)': 0.18121314992458898,
 'Oxygen saturation': -0.00797890030167256,
 'Total hardness': 0.6657262454837672,
 'Total dissolved salts': 0.05519675980558923,
 'Heptachloroepoxide': nan,
 'Heptachlor': nan,
 'Endosulfan (alpha)': nan,
 'Endosulfan (beta)': nan,
 'p,p-DDD': nan,
 'p,p-DDE': nan,
 'alpha-Hexachlorocyclohexane': nan,
 'beta-Hexaxchlorocyclohexane': nan,
 'Aldrin': nan,
 'Dieldrin': nan,
 'Endrin': nan,
 'Simazine': nan,
 'Metolachlor': nan,
 'Alachlor': nan,
 'Total nitrogen': -0.04642108309769432,
 'Tetrachloromethane': nan,
 'Cyanides': 0.2873131603614789,
 'Sulfides': nan,
 'Total organic carbon': 0.1443147005977184,
 'Dissolved silicates': -0.5494338864032309,
 'AGOC-3A': -0.005938025279960243,
 'Methylosmoline': 0.25453517491484445,
```

```
 'Chlorodinine': -0.8321079785699308,
 'Total dissolved phosphorus': 0.055125050071663295}
```

In [190…
```
high_corr = [key for key, value in corr.items() if (value > 0.5) or (value < -0.5)]
```

In [191…
```
high_corr
```

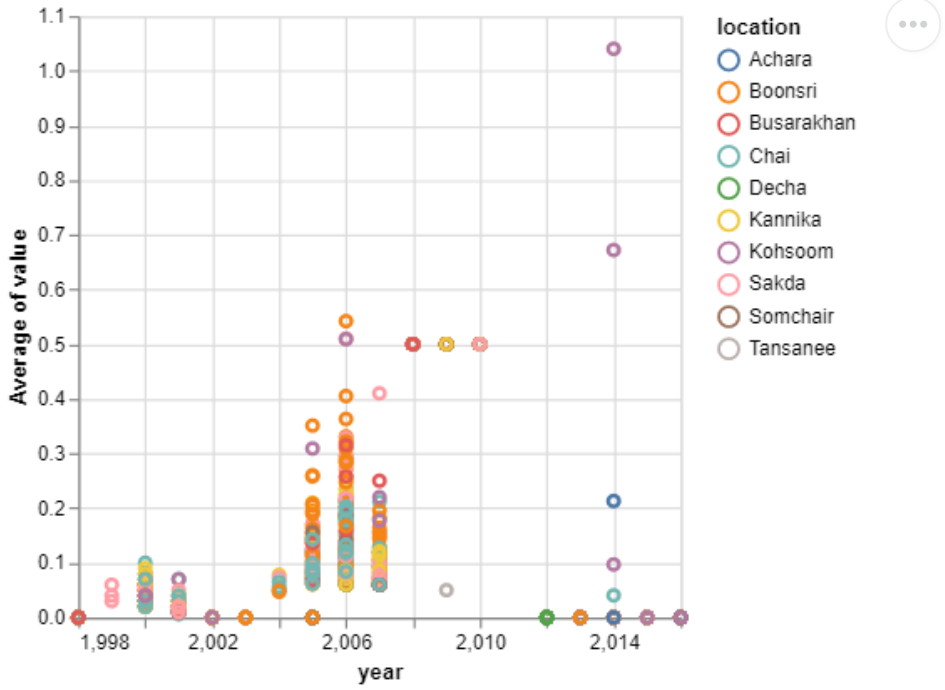
Out[191]:
```
['Atrazine',
 'gamma-Hexachlorocyclohexane',
 'Total hardness',
 'Dissolved silicates',
 'Chlorodinine']
```

In [192…
```
for a in high_corr:
    print(a,':', corr[a])
```

```
Atrazine : -0.8036555215998982
gamma-Hexachlorocyclohexane : -0.7528786840318084
Total hardness : 0.6657262454837672
Dissolved silicates : -0.5494338864032309
Chlorodinine : -0.8321079785699308
```

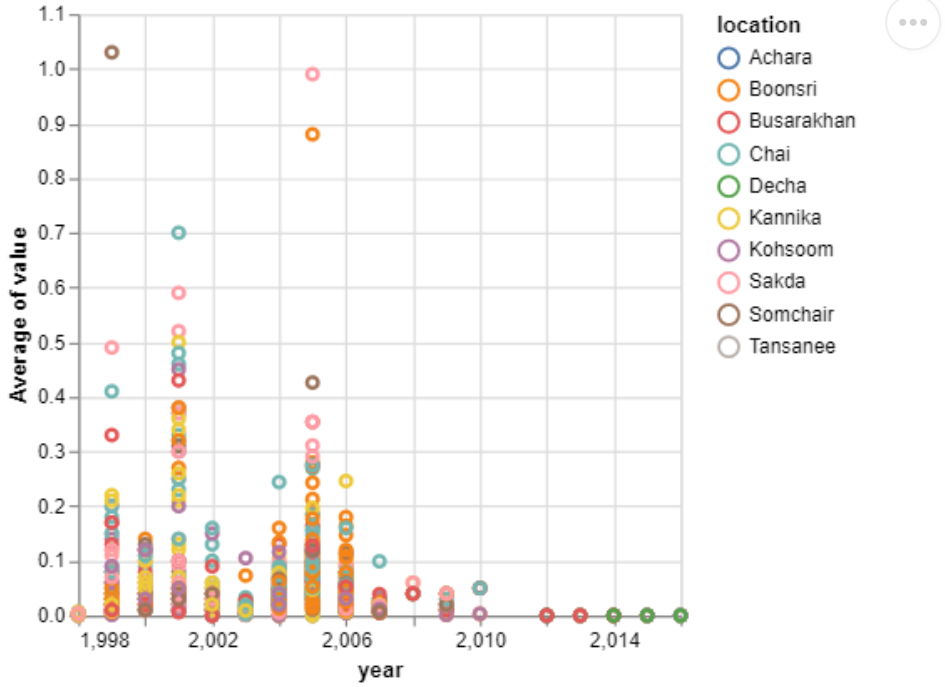
In [193…
```
chart = alt.Chart(df).mark_point().encode(
    alt.X(field='year', type='quantitative', scale=alt.Scale(zero=False)),
    alt.Y(field='value', aggregate='average', type='quantitative'),
    color='location:N',
tooltip=['value', 'sample date']).transform_filter(datum.measure=='Atrazine').prop
chart
```

Out[193]:



In [194…
```
chart = alt.Chart(df).mark_point().encode(
    alt.X(field='year', type='quantitative', scale=alt.Scale(zero=False)),
    alt.Y(field='value', aggregate='average'),
    color='location:N',
tooltip=['value', 'year']).transform_filter(datum.measure=='gamma-Hexachlorocycloh
chart
```

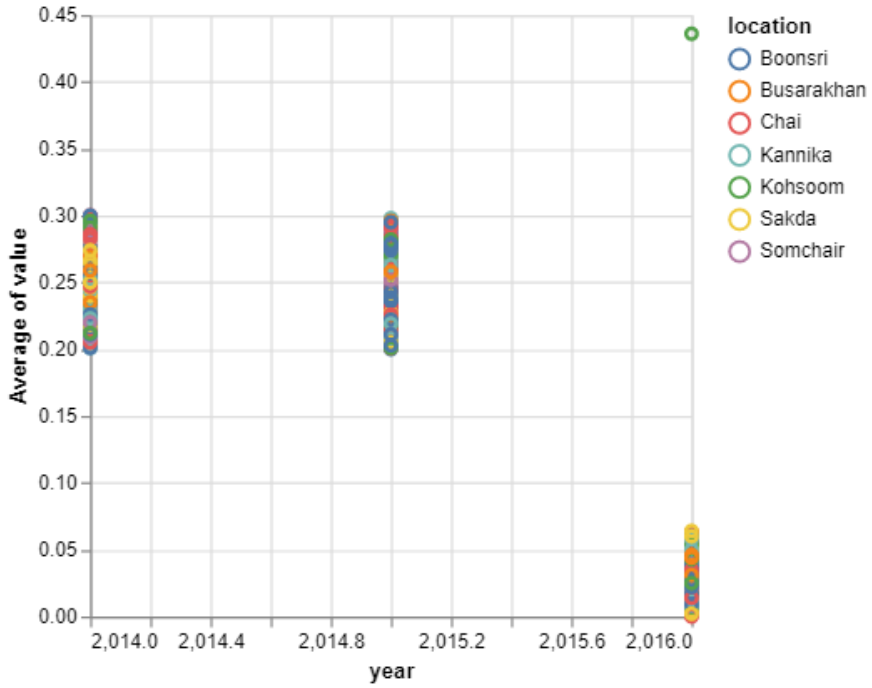
Out[194]:



In [195…

```
chart = alt.Chart(df).mark_point().encode(
    alt.X(field='year', type='quantitative', scale=alt.Scale(zero=False)),
    alt.Y(field='value', aggregate='average', type='quantitative'),
    color='location:N',
tooltip=['value', 'year']).transform_filter(datum.measure=='Chlorodinine').properti
chart
```

Out[195]:



In [196…

```
# All of these are unreliable for our analysis due to lack of observations + Lack o
```