VISUAL ANALYTICS

COURSEWORK - 2

M00977779 NIVEDITA KANNAN

Summarised findings list:

Q1. (i) Trends: The temporal trends of the top 5 measures were observed in all the locations mentioned in the dataset to identify a common pattern.

(ii) Anomalies: Iron is an anomaly as it peaks abnormally in 2003 compared to the other measures.

Q2. (i) Missing data: There are missing records in some locations, and not all locations have the same start time for data collection.

(ii) Change in collection frequency: Within the exact location and between two locations, the frequency of data collection is not the same.

(iii) Unrealistic values/Outliers: Some measures display a higher mean of 'value' than the rest.

(iv) Duplicate/Repeated values: Some dates have multiple values for the same measure and location, which plays an important role in reducing the reliability of the dataset.

In-depth explanation of code and goal achieved:

Packages and libraries imported:

Pandas, Altair, Seaborn, Matplotlib.pyplot, an iqr from scipy.stats.

Initial assessment and modification of the provided data:

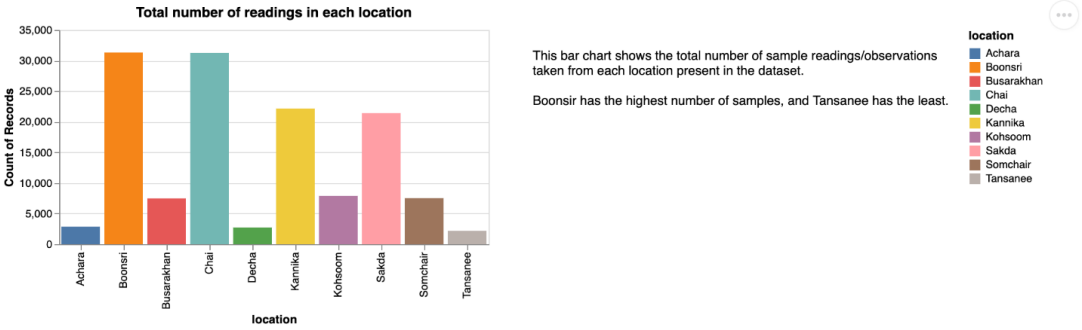
- Number of rows: 136824, number of columns: 5
- Columns in the dataset: *id, value, location, measure*
- Two additional columns were added: 'year' and 'month.'
- The dataset was checked for null values:

```
id           0
value        0
location     0
sample date  0
measure      0
dtype: int64
```

- The statistical description of the dataset was printed using "*describe()*'
- The number of rows in the dataset exceeds the maximum allowed (5000). This step enables working on large datasets.

  alt.data_transformers.disable_max_rows()
- Bar Chart:

  This chart displays the total number of observations recorded in each of the locations in the dataset, namely, '*Boonsri', 'Kannika', 'Chai', 'Kohsoom', 'Somchair', 'Sakdda', Busarakhan', 'Tansanee', 'Achara', 'Decha'*.



- Grouping:

  The dataset was grouped by *'measure'* names and the mean of the *'values'* was found and sorted.

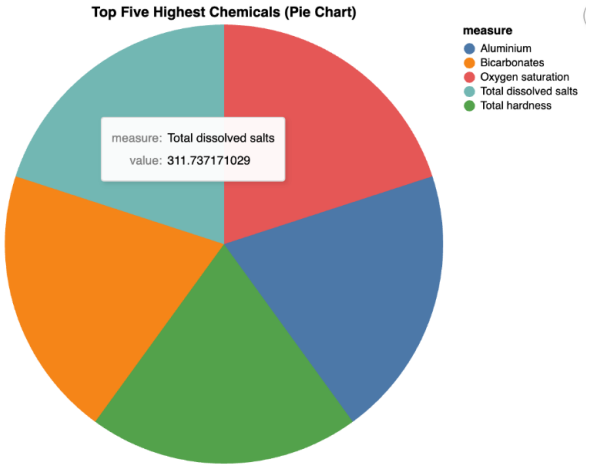
  The top five measures with the highest mean *'value'* are extracted and stored separately.

  Code Snippet:

```
# Calculate average values for each measure
average_values = data.groupby('measure')['value'].mean().reset_index()
# Sorting 'average values' in descending order
average_values_desc=average_values.sort_values(by='value',ascending=False)
print(average_values_desc)
# Isolating the top five highest mean 'value' chemicals/measures
top_5_measures = average_values_desc.head(5)
top_5_measures
#Storing the names of the measures with high mean 'value' in a variable
top_5_measures_names = top_5_measures['measure'].values
```

- Pie Chart:

  This chart displays the different mean ‘values’ for each of the top 5 measures



- Filtering:

  A new data frame is created to store the additional information about the top 5 measures. The analysis questions are mostly answered for these top 5 measures.

  Code Snippet:

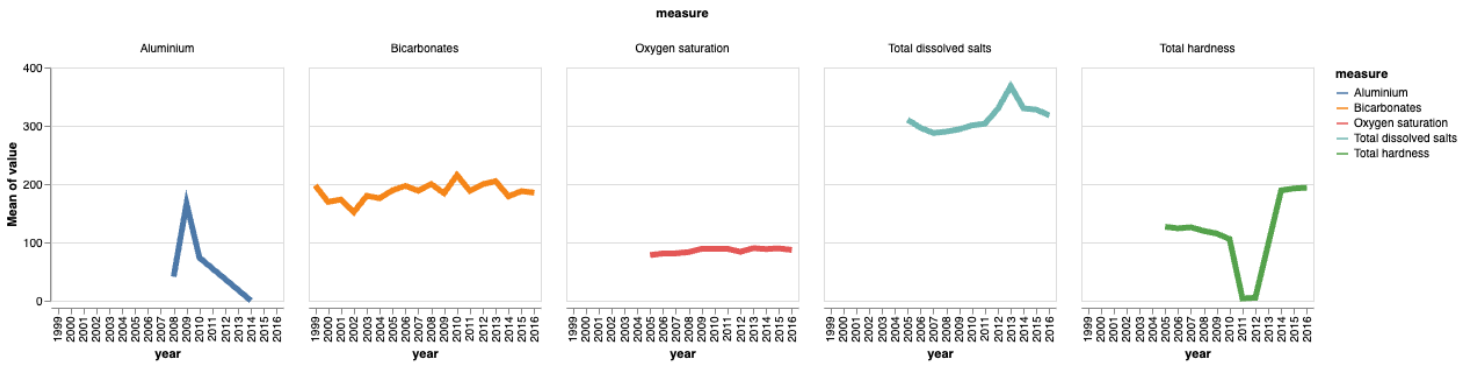
  filtered_data_top5 = data[data['measure'].isin(top_5_measures_names)]

Analysis Questions:

Q1:

(i) Trends

[What?] Temporal trend graphs for the top 5 measures were created and studied.



Select_Location_location Tansanee

[How?]A filter is applied to check the temporal trends for each location.

1. Aluminium: This measure has a steep positive slope from 2007 to 2009. The mean 'value' of aluminium had a short steep fall till 2010 and gradually moved further down until 2014.
2. Bicarbonates: Possesses a consistent rise and fall from 1999 to 2016.
3. Oxygen saturation: Maintains an approximately constant mean 'value' from 2005 to 2016.
4. Total dissolved salts: This measure initially decreases for a short period, peaks in 2013, and falls back to 2016.
5. Total hardness: The graph for this measure is quite peculiar. It depicts a huge variation in the mean 'value' of bicarbonates.

One more important observation is that locations such as Boonsri, Kohsoom, Tansanee, Achara, and Decha have no graphs for aluminium, indicating the lack of data for this measure. This is further proven in Q2.(i).

[This type of line graph is best suited to represent temporal trends as you can spot the rise and fall of the observed parameter over the years.]

[Visualisation features used: INTERACTION]

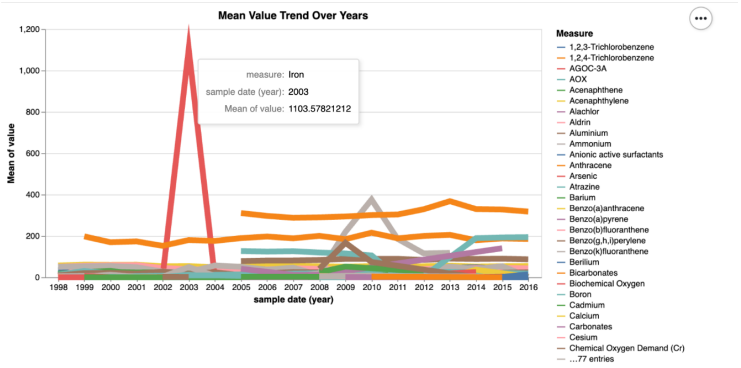
1. Filter dropdown: location
2. Tooltip

(ii) Anomalies:

[What?] Anomaly: Iron

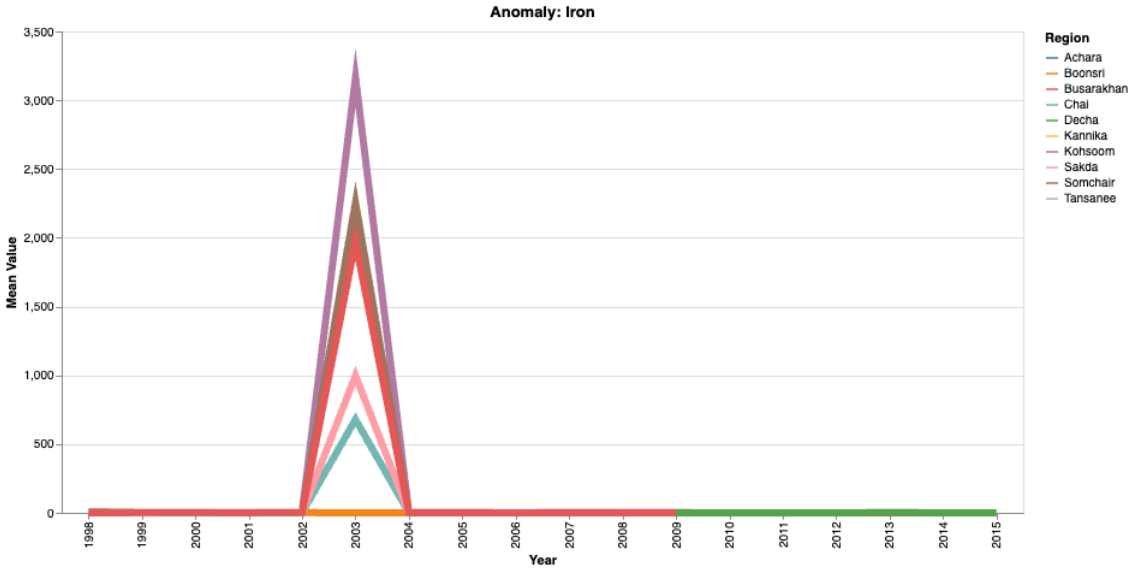
[How?] :

A line trend was created for the whole dataset to observe all the temporal trends possible with all the measures. In this first graph, it is evident that there is one particular temporal trend that stands out from the rest, the iron temporal trend. Hence, iron is isolated and observed in the second graph.



The tooltip can be used to view the mean 'value' for a specific measure in that specific year

The second temporal graph shows that in 2003, all the locations experienced a severe spike in the levels of aluminium.



Iron is the only measure showing a highly varied characteristic than the rest, making it an anomaly.

[Visualization features used:]

1. Tooltip

Q2.

(i) Missing data (ii) Change in data collection frequency

[What?]

1. **The data collection doesn't start in the same year for all locations.**
2. **Data collection is inconsistent i.e., some locations have year-wise gaps in data collection.**
3. **The frequency of data collection varies with the years. Different locations have different numbers of records.**

[How?]

Chart 1: A heat map of the total observations recorded in each location. The starting date for their records is 2009 for locations like Achara, Decha, and Tansanee, which is different from the others that started in 1998. The data isn't consistent for all the locations.

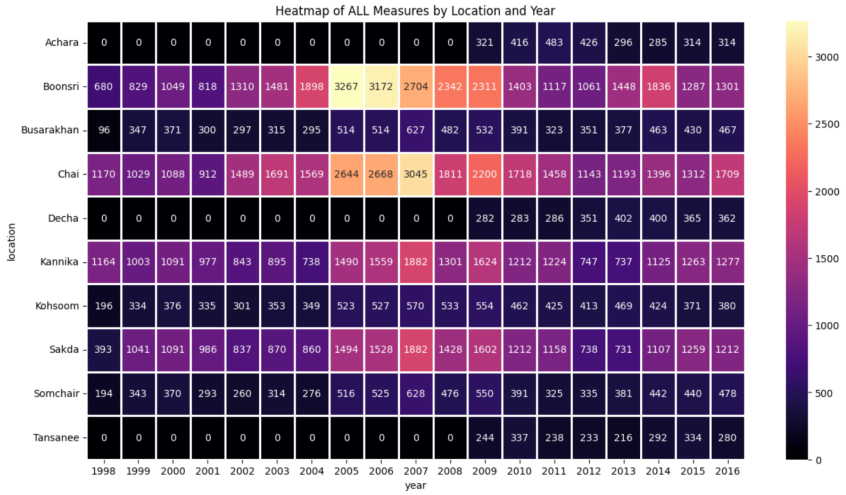


Chart 2: A heat map of total observations in the locations mentioned for the top 5 measures.

This chart is created to show missing observations in between a series of observations. The observations have been recorded in Achara since 2009, but data for 2010 to 2013 is missing.

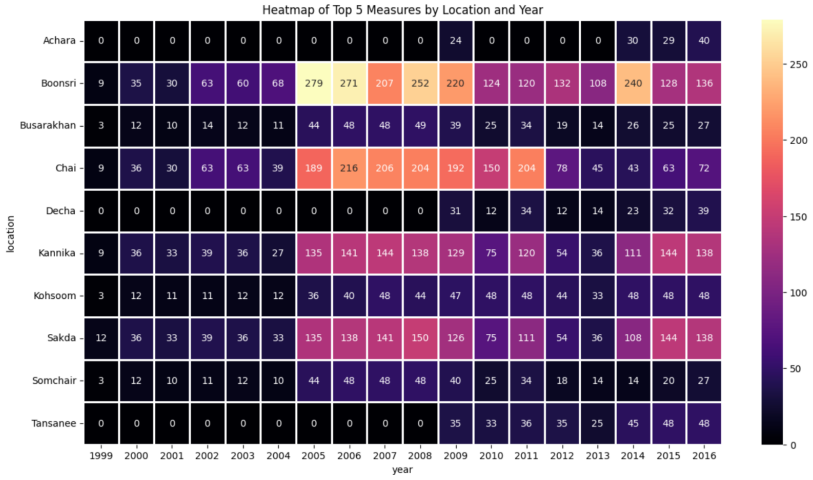
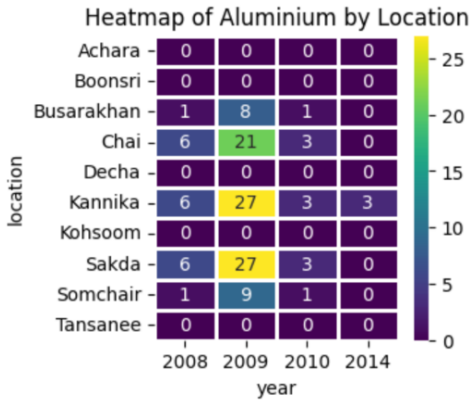


Chart 3: This heat map was created to prove the claim made in the first question that there is no data for “Aluminium” in Achara, Boonsri, Decha, Kohsoom, and Tansanee.

Even for locations with data for “aluminium”, the observations are collected at a highly varied frequency



The collection frequency is evidently inconsistent and highly varied, which makes the data unstable and unreliable.

No specific visualisation feature (multi-layer, chart concatenation, tooltip, interaction, etc.) is used in this finding. The input data is different for all three heatmaps.

(ii) Duplicate/Repeated values

[What?] This chart is created to observe if there are multiple values of observations for the same day.

The scatter plot is made for 2007, and the 'measure' of total dissolved salts is for easier analysis.

(Cropped images have been used.)

[How?]As you can see in the above graph, there are multiple circles (observations) for the same day. The tooltip can be used to view the values and the dates.

Example: There are three duplicates for the observations taken on the 24th of May, 2007, in Boonsri. The respective values of observation are 203, 216, and 227.

[Visualization features used:]

1. Tooltip
2. Facet

(iii) Outlier detection

[What?]

Identified outliers in the dataset:

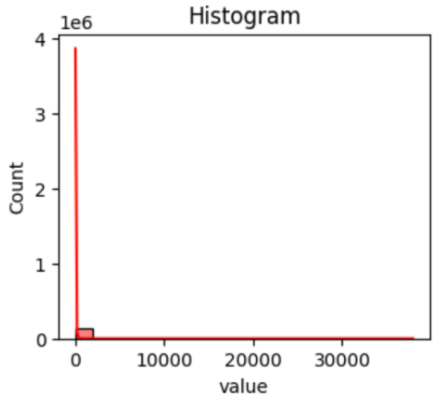
8	Aluminium
14	Barium
21	Bicarbonates
23	Boron
25	Calcium
26	Carbonates
28	Chemical Oxygen Demand (Cr)
30	Chlorides
52	Iron
56	Magnesium
68	Oxygen saturation
85	Sodium
87	Sulphates
89	Total coliforms
91	Total dissolved salts
93	Total hardness
98	Water temperature
99	Zinc

Name: measure, dtype: object

A histogram is plotted to check the normality of the dataset.

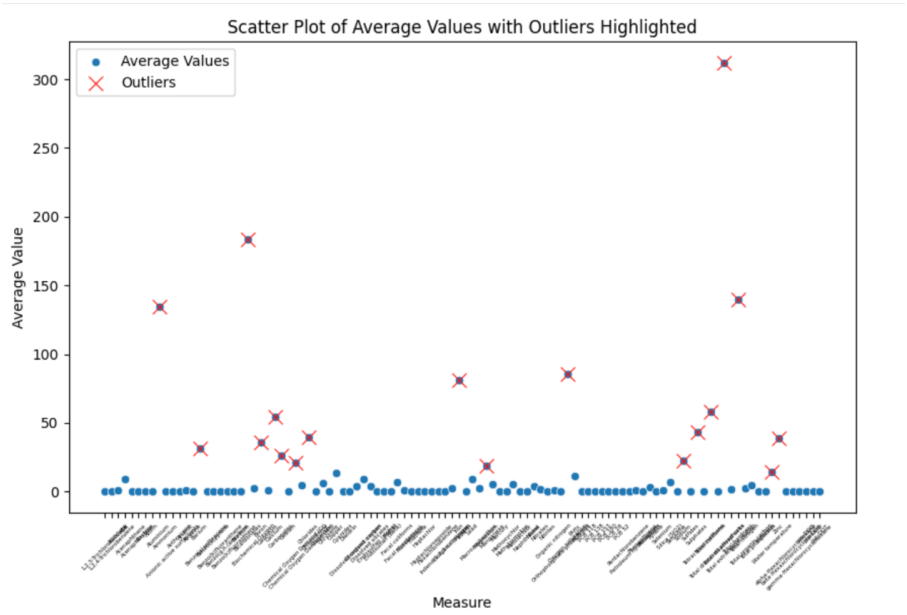
The graph shows that the data is right skewed; hence, the IQR method of outlier detection is used.

```
Text(0.5, 1.0, 'Histogram')
```



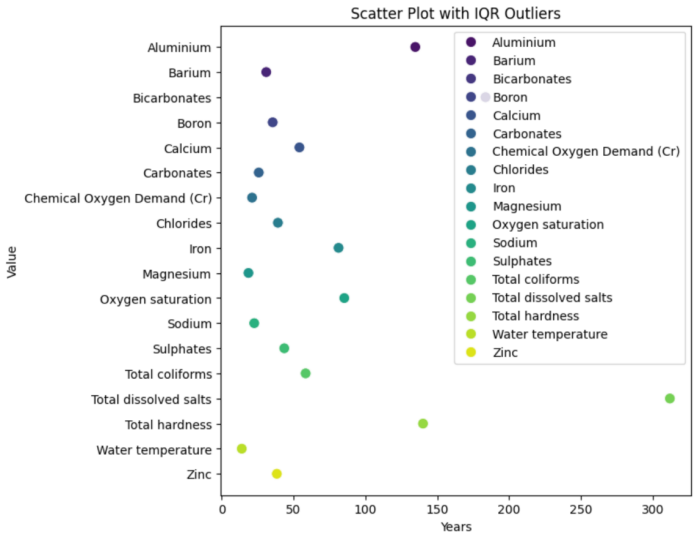
Above graph is a skewed graph, hence IQR, method is chosen for outlier detection.

Chart 1: This graph is plotted to visualise the outliers from the whole dataset.



[How?] The above graph shows the outliers marked in 'x' and the non-outliers in a circle.

Chart 2: This graph is plotted for the outliers alone.



[How?] This graph just represents the outliers alone.

The above two graphs depict which measures have the highest mean of 'value'. All the top 5 measures are also present in this list of outliers.